Parameter Estimation in Nonlinear Models of Biological Systems

W.R. Smith

Marine Ecology Laboratory
Bedford Institute of Oceanography
Ocean and Aquatic Sciences - Atlantic
Department of Fisheries and Oceans
Dartmouth, Nova Scotia
B2Y 4A2

March 1979

Fisheries and Marine Services Technical Report No. 889

Fisheries and Environment Canada
Fisheries and Marine Service

Pêches et Environnement Canada
Service des pêches et de la mer

Fisheries and Marine Service
Technical Reports

These reports contain scientific and technical information that represents an important contribution to existing knowledge but which for some reason may not be appropriate for primary scientific (i.e. *Journal*) publication. Technical Reports are directed primarily towards a world wide audience and have an international distribution. No restriction is placed on subject matter and the series reflects the broad interests and policies of the Fisheries and Marine Service, namely, fisheries management, technology and development, ocean sciences and aquatic environments relevant to Canada.

Technical Reports may be cited as full publications. The correct citation appears above the abstract of each report. Each report will be abstracted in *Aquatic Sciences and Fisheries Abstracts* and will be indexed annually in the Service's index to scientific and technical publications.

Numbers 1-456 in this series were issued as Technical Reports of the Fisheries Research Board of Canada. Numbers 457-714 were issued as Department of the Environment, Fisheries and Marine Service, Research and Development Directorate Technical Reports. The series name was changed with report number 715.

Details on the availability of Technical Reports in hard copy may be obtained from the issuing establishment indicated on the front cover.

Service des pêches et des sciences de la mer
Rapports techniques

Ces rapports contiennent des renseignements scientifiques et techniques qui constituent une contribution importante aux connaissances actuelles mais qui, pour une raison ou pour une autre, ne semblent pas appropriés pour la publication dans un journal scientifique. Il n'y a aucune restriction quant au sujet; de fait, la série reflète la vaste gamme des intérêts et des politiques du Service des pêches et de la mer, notamment gestion des pêches, techniques et développement, sciences océaniques et environnements aquatiques, au Canada.

Les Rapports techniques peuvent être considérés comme des publications complètes. Le titre exact paraîtra au haut du résumé de chaque rapport, qui sera publié dans la revue *Aquatic Sciences and Fisheries Abstracts* et qui figurera dans l'index annuel des publications scientifiques et techniques du Service.

Les numéros 1-456 de cette série ont été publiés à titre de Rapports techniques de l'Office des recherches sur les pêcheries du Canada. Les numéros 457-700, à titre de Rapports techniques de la Direction générale de la recherche et du développement, Service des pêches et de la mer, ministère de l'Environnement. Le nom de la série a été modifié à partir du numéro 701.

La page couverture porte le nom de l'établissement auteur ou l'on peut se procurer les rapports sous couverture cartonnée.

•Cover design by Christine Rusk

Fisheries and Marine Service

Technical Report No. 889

March 1979

PARAMETER ESTIMATION IN NONLINEAR MODELS OF BIOLOGICAL SYSTEMS

by

William R. Smith

Department of Mathematics, Department of Physiology and Biophysics

Dalhousie University, Halifax, N.S. B3H 4H8 Canada

This is the sixty-fifth Technical Report from the Marine Ecology Laboratory, Bedford Institute of Oceanography, Dartmouth, Nova Scotia, and was prepared for the Marine Ecology Laboratory under personal services contract OSC78-00154.

© Minister of Supply and Services Canada 1979

Cat. no. Fs 97-6/889 ISSN 0701-7626

TABLE OF CONTENTS

		Page
1.	THE MATHEMATICAL AND STATISTICAL MODELS	1
1.1	Introduction	1
1.2	Model-Fitting versus Curve-Fitting	4
1.3	Linear versus Nonlinear Mathematical Models	9
1.4	The Statistical Model	13
2.	ESTIMATION CRITERIA FOR THE PARAMETERS	16
2.1	Maximum Likelihood Estimation	16
2.2	Bayesian Estimation	21
2.3	Least-Squares Estimation	23
2.4	Linearizing Transformations	25
3.	NUMERICAL COMPUTATION OF THE PARAMETER ESTIMATES ...	28
3.1	Introduction	29
3.2	Search Methods	31
3.3	Derivative Methods	32
3.4	Differential Equation Methods	38
4.	THE PRECISION OF THE PARAMETER ESTIMATES	40
4.1	The Sampling Distribution of the Estimator ..	41
4.2	Approximate Computation of the Covariance Matrix of the Estimates	42
4.3	Approximate Joint Confidence Regions	45
4.4	Principal Components	48
4.5	Confidence Intervals	49
4.6	Exact Computation of the Sampling Distribution	50
5.	FURTHER IMPLICATIONS OF MODEL-FITTING	52
5.1	Examination of the Residuals	52
5.2	The Precision of Model Predictions	54
5.3	Experimental Design	55
6.	SOME SPECIAL SITUATIONS	60
6.1	Separable Least-Squares Problems	60
6.2	Differential Equation and Other Implicit Models	62
6.3	Linear Compartmental Models	65
6.4	Independent Variables Subject to Error	72
7.	A WORKED EXAMPLE	74
8.	SUMMARY OF THE PROCEDURE	80
REFERENCES		83

ABSTRACT

Smith, W.R. 1979. Parameter estimation in nonlinear models of biological systems. Fish. Mar. Serv. Tech. Rept. 889: 90 p.

The problem of estimating numerical values of parameters appearing in mathematical models of biological systems is discussed in detail. Particular emphasis is placed on the computational aspects of the problem and the statistical interpretation of the results. A concise summary of a recommended practical procedure is given and this is illustrated in the context of a worked example.

SOMMAIRE

Smith, W.R. 1979. Parameter estimation in nonlinear models of biological systems. Fish. Mar. Serv. Tech. Rept. 889: 90 p.

On discute en détail du problème de l'évaluation des valeurs numériques des paramètres se présentant dans les modèles mathématiques de systèmes biologiques. On met un accent particulier sur les aspects de calcul du problème et sur l'interprétation statistique des résultats. On donne un résumé concis d'une procédure pratique recommandée et on fait la démonstration de cette procédure au moyen d'un exemple.

CHAPTER 1

THE MATHEMATICAL AND STATISTICAL MODELS

1.1 INTRODUCTION

We will be discussing herein the problem of estimating the values of a set of parameters $\underline{p}^{\dagger}$ from observations of some response function $\underline{y}$ which depends on some independent variable $\underline{x}$, in addition to the parameters. The parameter and independent variable dependence will be assumed to be described by some postulated *mathematical model* for the response denoted by

$$y = f(x,\underline{p}) \tag{1.1}$$

Typically, a set of values of y (observations) is known corresponding to a set of x values, yielding the *observation set* $\{(x_i,y_i)|i = 1,2,\ldots n\}$ (n being the number of observations). If the x_i and y_i values are known exactly (without error), then the problem of the determination of the function f and parameters p is called an *approximation* problem. We are not concerned with this case, but with the case in which x_i and/or y_i contain errors. The problem is hence to determine the parameters from the set of *observational equations*

$$y_i = f(x_i + e_i',\underline{p}) + e_i \ ; \tag{1.2}$$

e_i and e_i' are the errors of y_i and x_i, respectively. One of the most commonly assumed situations in parameter estimation problems occurs when

---

† In this manual, all vectors are column vectors and are denoted by underlining. Thus $\underline{p}$ is the column vector $(p_1,p_2,\ldots p_m)^T$. Superscript T denotes a transposed quantity. We assume that $m < n$, where n is the number of observations.

x_i is known exactly (i.e., $e_i' \equiv 0$) and only the dependent variable is subject to error. Thus Equation 1.2 becomes

$$y_i = f(x_i, \underline{p}) + e_i \ . \tag{1.3}$$

This assumption is not always valid in practice. However, we will usually assume its validity in what follows. We discuss the case of the errors in the independent variable in Section 6.4.

The errors e_i in Equation 1.3 are in general of two types. The first is due to the possible inappropriateness of the model function f, and the second is due to errors of measurement of the response function values y_i. We will usually assume only the second type occurs in a given practical situation; i.e. we assume we are always using the correct mathematical model for the situation. For some ways of discriminating between alternative models, see for example Section 10.5 of Bard (1974).

We also encounter the case where there is more than one independent variable. Then x in Equation 1.3 becomes a vector $\underline{x}$ and Equation 1.3 becomes

$$y_i = f(\underline{x}_i, \underline{p}) + e_i \ . \tag{1.4}$$

Finally, the response function may also be a vector, giving

$$\underline{y}_i = f(\underline{x}_i, \underline{p}) + \underline{e}_i \ . \tag{1.5}$$

Equation 1.5 is appropriate when one measures several different properties (y-values) of the same biological system. In most of what follows, we will

be considering situations appropriate to the use of Equation 1.3 or 1.4, and will indicate the extensions required to cover other cases.

We assume that the function(s) f in Equations 1.1 to 1.5 are known at the outset. We will discuss this assumption briefly in the next section. The model function f may be given *explicitly* in analytical form - for example

$$f(x,\underline{p}) = p_1 \exp(p_2 x) \tag{1.6}$$

is a model applicable to many experimental situations, and we will often use it for illustrative purposes later in this manual. Sometimes the model function is only known *implicitly*. For example, it may be the solution of an ordinary differential equation

$$\psi\left(x,\underline{p},f,\frac{df}{dx},\frac{d^2f}{dx^2},\ldots,\frac{d^rf}{dx^r}\right) = 0 . \tag{1.7}$$

If ψ is linear in the derivatives as well as in x and f, then we may typically recover the solution of Equation 1.7 analytically in the form

$$f(x,\underline{p}) = \sum_{i=1}^{r} \alpha_i \exp(\beta_i x) \tag{1.8}$$

where the parameter set $\underline{p}$ has been subdivided into the two parameter sets $\underline{\alpha}$ and $\underline{\beta}$. Equation 1.8 is called a *compartmental model* and is widely used in many areas of biology. (We discuss compartmental models in detail in Section 6.3.) In many situations, however, we cannot solve Equation 1.7 in closed form and must integrate the equations numerically. We will describe the treatment of these *implicit* models in Section 6.2. The main point to emphasize here is that all we require of our model is the ability to compute

(numerically if necessary) corresponding values of f for a given set of $\underline{p}$ and x.

1.2 Model-Fitting versus Curve-Fitting

Here we discuss the choice of the function f in Equation 1.1. The remainder of this manual deals with how to proceed for a given form of f, but as every scientist knows, parameter estimation is usually part of the process of constructing a *mathematical model* of some observed system. Mathematical modelling can be said to consist of two main stages. The first of these is the construction of an appropriate model function f, and the second stage is the estimation of the parameters in f from a set of observational data on the system.

One of the most important goals of the mathematical modelling process is:

(1) Prediction: To predict the response function y for values of x not in the existing observation set.

This goal, in turn, can be subdivided into two cases:

(1a) Interpolation: To predict y at values of x that lie within the range of measured x_i values (or at x values not too far outside that range).

(1b) Extrapolation: To predict y at values of x far removed from the range of measured x_i values.

This subdivision is important because functions f appropriate to each purpose will often be different. We call functions f appropriate to Goal (1a) *interpolation models* and those appropriate to Goal (1b) *extrapolation models*. The parameters in a given extrapolation model are usually determined in practice by using some observation set $\{(x_i,y_i)\}$; that is, by using f as

an interpolation model for some observation set.

It should be self-evident that extrapolation models are to be desired since they can also be used for interpolation, but the reverse is not generally true. If the process being observed is understood in terms of some known or assumed *fundamental laws*, then this can produce an extrapolation model. Difficulties arise in practice because of the complexity of most biological processes, and the absence of many fundamental laws.

In many situations arising in chemistry and physics, one can often construct extrapolation models. For example, an extrapolation model for radioactive decay is Equation 1.6. In the social sciences and economics, on the other hand, there are often no fundamental laws available (or at least few seem to have been found). Then we must of necessity construct arbitrary interpolation models and hope that they can be used for extrapolation. Biological systems probably lie somewhere in between the two extremes, and in Table 1.1 we illustrate this.

Table 1.1. Types of Models Arising in Various Disciplines

Extrapolation Models	?	Interpolation Models
Chemistry	Biology	Social Sciences
Physics		Economics

The dangers of using interpolation models for extrapolation are amply illustrated by considering Equation 1.3 when all the measurements are exact, i.e., the approximation problem. For example, suppose the process being observed is actually described by

$$y = e^x \tag{1.9}$$

and we measure x at n values of x in the interval $[0,\pi]$. A widely used interpolation model is the polynomial function,

$$f(x,p) = \sum_{i=0}^{r} p_i \, x^i \tag{1.10}$$

and we can determine the parameters $\underline{p}$ so that $f(x,\underline{p})$ agrees exactly with the observed y_i values by using n+1 parameters. Even if we insist that the polynomial f should accurately predict y at *every* x in $[0,\pi]$, in the sense that

$$\max_{x\epsilon[0,\pi]} |y - f(x,\underline{p})| < \varepsilon \tag{1.11}$$

where ε is an arbitrarily small positive number, then we can still succeed. This is a consequence of the Weierstrass approximation theorem of analysis (Kreider *et al.*, 1966) which states roughly that an arbitrary continuous function can be approximated arbitrarily closely (in the sense of Equation 1.1) over a closed and bounded interval by a polynomial of sufficiently high degree. Having found suitable interpolation models for the above problem, we can then enquire of their possible use for purposes of extrapolation outside the interval $[0,\pi]$. It is clear, however, that a polynomial approximation which agrees well with e^x over $[0,\pi]$ must ultimately fail outside this interval.

Having emphasized that good interpolation models of an inappropriate functional form can be unsuitable for extrapolation, we also wish to emphasize that interpolation models are not unique. For the example above, we would also choose an approximating function of the form

$$f(x,p) = \sum_{i=0}^{r} p_i \cos(ix) \ . \qquad (1.12)$$

Provided r is large enough, we can still satisfy inequality 1.11. In Equation 1.10 we have constructed the model from the approximating functions $\{x^i | i=0,1,2,\ldots\}$ and in Equation 1.12 we have used the set $\{\cos(ix) | i=0,1,2,\ldots\}$. The Weierstrass approximation theorem mentioned above only states that the set of approximating functions must be a *basis set* for the vector space of continuous functions defined on $[0,\pi]$ in order for inequality 1.11 to hold.

In constructing an interpolation model when we have no guide as to the appropriate choice of the model function f, we usually choose it so that its properties make the parameter estimation stage 'simple' in some way. This 'simplicity' has both computational and statistical implications, which we explore in Section 1.3. 'Simplicity' is typically achieved by the use of a polynomial function in the case of a single independent variable x and a linear function of $\underline{x}$ in the case of several independent variables. Parameter estimation techniques for these cases are extensively discussed in many statistics texts and computer packages are available for their implementation. The IMSL package 'LLSQF' is probably the best of these, implementing the algorithms in Lawson and Hanson's (1974) text.

We will henceforth call *model-fitting* the process of estimating the parameters in an extrapolation model whose functional form is chosen with some regard to the underlying biology, and *curvefitting* the corresponding process when an interpolation model whose functional form is arbitrarily chosen is used. The two problems have certain general similarities but important practical differences, which we will explore in Section 1.3.

A second important goal of mathematical modelling is often:

(2) Insight: To determine estimates of the parameters, denoted by $\hat{\underline{p}}$, from the given observational data with the objective of using these estimates in models of other related biological systems. This implies also a determination of the precision of the estimates $\hat{\underline{p}}$.

If the model function f is constructed on the basis of some assumed fundamental laws (an extrapolation model) then $\underline{p}$ usually represents parameters in these underlying laws. These parameters may then be used in constructing mathematical models of biological processes different from the one of current interest. The accuracy with which the parameters are known when estimated from the given set of observational data is then of interest. We call this the *precision* of the estimates. In curve-fitting, the precision of the parameter estimates for the model is often not of any great interest, since the prediction of y values is usually the principal goal [Goal (1)].

As an illustration, suppose the biological system comprises, among other things, the biochemical reaction

$$A \rightarrow B \ . \tag{1.13}$$

The chemical law of mass-action states that the rate of change of the concentration of B in the system is described by the differential equation

$$\frac{dB}{dt} = k\, A^{\alpha}\, B^{\beta} \ . \tag{1.14}$$

If the observations y of the system depend in some way on the concentration of B, then the parameters $\{k,\alpha,\beta\}$ in Equation 1.14 are a subset of $\underline{p}$ which determines y. In a sense, we can think of Equation 1.14 as a 'submodel' for y. The parameters $\{k,\alpha,\beta\}$ are of interest in their own right, since for all other biological systems in which the chemical reaction of Equation 1.13 occurs, their values remain unchanged. When we have in mind using the parameters in these other situations, we are interested not only in estimation of their values, but in the precision of these estimates. The determination of the precision of the parameter estimates in a parameter estimation problem is discussed in detail in Chapter 4.

To summarize, there are three important computational aspects of model-fitting. The first is the actual numerical estimation of the parameters occurring in the function f. We discuss this in Chapter 3. The second is the determination of the precision of the parameter estimates computed in the first stage. We discuss this problem in Chapter 4. The third is determining the precision of values of y predicted from the model function f at values of x not previously considered. We discuss this problem in Section 5.2. In curve-fitting, only the first and third aspects are usually of direct interest.

1.3 Linear versus Nonlinear Mathematical Models

The actual techniques of parameter estimation to be discussed in the remainder of this manual can be used both for curve-fitting and for model-fitting. As discussed in Section 1.2, in curve-fitting, the model function f is usually chosen so as to make the parameter estimation problem 'simple.' We wish to explore the meaning of this simplicity here. It results from two separate factors, one related to the functional form

of f and the other related to the statistical criterion used to estimate the parameters $\underline{p}$ (see Section 1.4).

The first factor is reflected in the fact that the model function f is chosen so that it is *linear in the parameters*. *Linear models* are model functions of the form

$$f(\underline{x},\underline{p}) = \sum_{j=1}^{m} p_j \phi_j(\underline{x}) \ . \tag{1.15}$$

It is important to note that the linearity refers to the parameters $\underline{p}$ and not to the independent variables $\underline{x}$. Specific examples of linear models are:

$$f_1(x,\underline{p}) = p_1 + p_2 x + p_3 x^2 \tag{1.16}$$

$$f_2(x,\underline{p}) = p_1 + p_2 \sin x + p_3 \cos x \tag{1.17}$$

$$f_3(\underline{x},\underline{p}) = p_1 + p_2 x_1 + p_3 x_2 + p_4 x_3. \tag{1.18}$$

Equations 1.16 and 1.17 are nonlinear in x while Equation 1.18 is linear in $\underline{x}$, but all three are called linear models in the context of parameter estimation.

The second factor contributing to 'simplicity' is that the appropriate criterion for estimating $\underline{p}$ is the minimization of a quadratic function of the differences between the model predictions and the observations over the observation set $\{(x_i,y_i)\}$. A typical example is the least-squares criterion (Section 2.3), in which $\hat{\underline{p}}$ is determined by solving the minimization problem

$$\min_{\underline{p}} \{S(\underline{p}) = \sum_{i=1}^{n} [y_i \quad f(x_i,\underline{p})]^2\}. \tag{1.19}$$

This criterion dates back to Gauss (1809). Its statistical appropriateness depends on the properties of the probability density functions of the measurement errors in Equation 1.3 (the 'statistical model' see Section 1.4) and we will discuss this and other criteria for estimating $\underline{p}$ in Chapter 2.

The 'simplicity' of Equations 1.15 and 1.19 refers to the fact that the first-order conditions for the minimization of $S(\underline{p})$ in Equation 1.19, given by

$$\frac{\partial S}{\partial p_j} = 0; \quad j=1,2,\ldots m \tag{1.20}$$

where m is the number of parameters in Equation 1.15, yield a set of *linear* algebraic equations for the estimates $\hat{\underline{p}}$. Equations 1.20 are in general called the *normal equations* and in the present case of a linear model function and the quadratic minimization criterion of Equation 1.19 are given by

$$\sum_{k=1}^{m} \hat{p}_k \sum_{i=1}^{n} \phi_j(x_i)\phi_k(x_i) = \sum_{i=1}^{n} \phi_j(x_i)y_i; \quad j=1,2,\ldots m \ . \tag{1.21}$$

It is readily established that the solution $\hat{\underline{p}}$ of Equations 1.21 provides the unique minimum of $S(\underline{p})$, provided the functions ϕ_j are linearly independent and the x_i are distinct. Also, linear equations are straightforward to solve, at least on an electronic computer. We remark, however, that Equations 1.21 are of some nontrivial numerical interest, since the coefficient matrix is often ill-conditioned (see, for example, Lawson and Hanson, 1974). Thus, for linear models and quadratic minimization criteria, the solution to the parameter estimation problem exists, is usually unique, and

is relatively easy to compute. It also turns out that the determination of the precision of the resulting parameter estimates and the precision of predictions of y using these estimates is relatively straightforward (Draper and Smith, 1966).

If either f is nonlinear in the parameters or a nonquadratic minimization criterion is used, then the normal equations will be *nonlinear*. Thus, for example, if Equation 1.19 remains unchanged and the model function f is nonlinear, the resulting nonlinear normal equations are

$$\sum_{i=1}^{n} f(x_i,\underline{p})\left(\frac{\partial f}{\partial p_j}\right) = \sum_{i=1}^{n}\left(\frac{\partial f}{\partial p_j}\right) y_i;\ j=1,2,\ldots m \qquad (1.22)$$

where the derivatives are evaluated at x_i. The solution of Equations 1.22 is a nontrivial numerical problem. It is rarely easy to establish that solutions are unique, and solutions are sometimes difficult to compute, even using an electronic computer (see Chapter 3). We will also see in Chapters 4 and 5 that when the model is nonlinear the problem of determining the precision of the parameter estimates and the precision of the predictions of y using these estimates is also beset with difficulties.

It is probably true that the vast majority of extrapolation models used in science which are based even partly on known or assumed 'fundamental laws' fail to satisfy the assumption of linearity of the model function in the parameters. For example, for the radioactive decay model of Equation 1.6, the least-squares criterion for estimating $\underline{p}$ leads to the minimization problem†

$$\min_{\underline{p}}\{S(\underline{p}) = \sum_{i=1}^{n}[(y_i - \alpha\exp(\beta x_i)]^2\}. \qquad (1.23)$$

† The astute reader will recognize that Equation 1.6 can be transformed into a linear model. This will be discussed in Section 2.4. Most nonlinear models cannot be linearized in this way.

This results in the nonlinear normal equations

$$\begin{aligned} \alpha \sum_{i=1}^{n} \exp(2\beta x_i) &= \sum_{i=1}^{n} y_i \exp(\beta x_i) \\ \alpha \sum_{i=1}^{n} x_i \exp(2\beta x_i) &= \sum_{i=1}^{n} x_i y_i \exp(\beta x_i). \end{aligned} \tag{1.24}$$

The general community of 'statisticians' is no different from the general community of 'pure mathematicians' in the fact that linearity often plays a large role in the problems considered. Thus, most statistical research in the past concerning parameter estimation has concentrated on linear models. However, due primarily to the advent of electronic computers, nonlinear parameter estimation has recently become an active area of mathematical and statistical research, and several recent textbooks and articles are devoted to a general discussion of the problem (Hartley and Booker, 1963; Wollberg, 1967; Chambers, 1973, 1977; Bard, 1974; Seinfeld and Lapidus, 1974; Gallant, 1975; Swartz and Bremermann, 1975; Mikhail, 1976; Jennrich and Ralston, 1979).†

1.4 The Statistical Model

Equation 1.3 actually consists of two terms. The first is the mathematical model $f(x,\underline{p})$ and the second is e_i, the error term. The collection of known (or as is more usual, assumed) characteristics of the errors e_i is called the *statistical model* for the error in the measured response y_i. As discussed in Section 1.1, this actually contains two terms - systematic error (in the model) and measurement error, although we usually assume e_i

† Note added in proof: Volume 8, No. 1 (1977) of the journal *Math. Operationsforsch. Statist., Ser. Statistics* contains a number of relevant articles concerning parameter estimation.

represents only the latter. We will see in Chapter 2 that the criterion we use to estimate the parameters depends exclusively on the statistical model.

The most general statistical model is one in which the errors are random variables having some joint probability density function $\psi(\underline{\varepsilon},\underline{\theta})$, where $\underline{\theta}$ is a set of parameters of the density function. ψ gives the probability of observing errors in a neighbourhood of volume $\delta\underline{\varepsilon}$ of a given $\underline{e}$ as being proportional to $\psi(\underline{e},\underline{\theta})\delta\underline{\varepsilon}$.

The form of statistical model most often used in practice assumes ψ is a normal probability density. Sometimes this is done on the basis of knowledge of the measurement process, but often it is simply assumed at the outset. Draper and Smith (1966) on p. 17 argue that there is a strong tendency in many real situations for errors to be normally distributed. This is true when the error term is a sum of errors from several sources, the result tending to normality by the Central Limit Theorem.

In many cases, however, the errors cannot arise from a normal distribution. For example, if y_i must be strictly positive due to the nature of the quantity, this restricts e_i to be bounded below by $-y_i$. Ideally, of course, we would like our estimating criterion used to determine $\hat{\underline{p}}$ to be relatively insensitive to the departure of the errors from normality. This degree of insensitivity of an estimator is sometimes called its 'robustness,' and is a general notion of considerable current interest in statistics, due largely to the work of Tukey (1975) and others (Huber, 1973; Hampel, 1973; Andrews, 1974; Lide and Paul, 1974; and Rey, 1978).

Some workers, in fact, argue strongly against the assumption that errors are normal (for example, Rice and White, 1964). However, estimators based on this assumption are apparently sufficiently robust that they never

produce grossly misleading results (Endrenyi and Kwong, 1972; Endrenyi, 1976). In any event, there are available techniques for testing the assumption of normally distributed measurement errors after the parameters have been computed (Section 5.1).

The two commonest forms of error behavior which occur (or are assumed) in practice are the following:

Type A: e_i are independent random variables following a normal probability distribution with mean 0 and constant variance σ^2. Usually σ^2 is unknown in advance of the experiments.

Type B: e_i are independent random variables following a normal probability distribution with means 0 and variances $\sigma^2(y_i^t)^2$ where y_i^t is the 'true' value of the response y_i and σ^2 is usually unknown.

Type A behavior is called the case of *constant absolute error variance* and Type B behavior is called the case of *constant relative error variance*. We will briefly mention more general types of normal error distributions in Chapter 2, but we emphasize Type A and Type B behavior in the rest of this manual. In later sections, we will see how to estimate the unknown quantity σ^2 (Section 2.1) and how to check which of Type A or Type B behavior is actually present (Section 5.1).

CHAPTER 2

ESTIMATION CRITERIA FOR THE PARAMETERS

2.1 Maximum Likelihood Estimation

We consider Equation 1.3, and wish to obtain an estimate $\hat{\underline{p}}$ of the 'true' parameters $\underline{p}$ from the observational equations

$$y_i = f(x_i,p) + e_i;\ i=1,2,\ldots,n. \tag{2.1}$$

The errors $\{e_i\}$ are drawn from some joint probability density function which we denote by $\psi(\underline{\varepsilon},\underline{\theta})$, where $\underline{\theta}$ are parameters of the density function (the 'statistical model' of Section 1.4).

The probability of observing a set of errors in some neighborhood of a given $\underline{\varepsilon}$ is proportional to $\psi(\underline{\varepsilon},\underline{\theta})\delta\underline{\varepsilon}$, where $\delta\underline{\varepsilon}$ denotes the volume element. The *likelihood function* results from replacing $\underline{\varepsilon}$ in the density function by the actually observed e_i from Equation 2.1. The likelihood function depends on the true model parameters $\underline{p}$ and the error density parameters $\underline{\theta}$ and is defined by

$$L(\underline{p},\underline{\theta}) = \psi(\underline{y} - \underline{f}\ (\underline{x},\underline{p}),\theta) \tag{2.2}$$

where

$$\underline{y} = (y_1,y_2,\ldots,y_n)^T \tag{2.3}$$

and we have used the notation $\underline{f}(\underline{x},\underline{p})$ to denote

$$\underline{f}(\underline{x},\underline{p}) = [f(x_1,\underline{p}),f(x_2,\underline{p}),\ldots,f(x_2,\underline{p})]^T.$$

The *maximum likelihood estimates* (MLE) of $\underline{p}$ are obtained by maximizing the likelihood function of Equation 2.2. Thus, the MLE parameter estimates are obtained by solving the problem

$$\max_{\underline{p}}\{S^*(\underline{p}) = L(\underline{p},\underline{\theta})\} \tag{2.4}$$

or equivalently, by solving

$$\min_{\underline{p}}\{S(\underline{p}) = -\log L(\underline{p},\underline{\theta})\}. \tag{2.5}$$

In an intuitive way, the MLE parameters $\hat{\underline{p}}$ are those which cause the observed values of e_i in Equation 2.1 to maximize the corresponding observed value of the joint density function of the measurement errors.

The implementation of the MLE procedure is usually achieved either by solving (numerically) the minimization problem of Equation 2.5 or by solving the resulting first-order conditions for the m parameters:

$$-\frac{\partial \log L}{\partial p_j} = 0;\ j=1,2,\ldots,m. \tag{2.6}$$

Equations 2.6 are called the *likelihood equations* for the parameters, and are in general a set of nonlinear equations.

Some special forms of $L(\underline{p},\underline{\theta})$ occur sufficiently often in practice that they are worthwhile considering at this point. One such case occurs when the errors are independent with probability density functions $\psi_i(\underline{\epsilon},\underline{\theta}_i)$ yielding

$$L(\underline{p},\underline{\theta}) = \prod_{i=1}^{n} \psi_i(e_i,\underline{\theta}) \tag{2.7}$$

Then Equations 2.5 become

$$\min_{\underline{p}}\{S(\underline{p}) = -\sum_{i=1}^{n} \log \psi_i(e_i,\theta_i)\}. \tag{2.8}$$

Another special case occurs when the errors have a *multivariate normal distribution*,

$$\psi(\underline{\varepsilon},\underline{\theta}) = \frac{(2\pi)^{-n/2}}{[\det(\underline{V})]^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}\,\underline{\varepsilon}^T\underline{V}^{-1}\underline{\varepsilon}\right) \tag{2.9}$$

where $\underline{V}$ is the $(n\times n)$ covariance matrix. (In the above case $\underline{\theta}$ represents the entries of $\underline{V}$). Then we have

$$-\log L(\underline{p},\underline{\theta}) = \frac{n}{2}\log 2\pi + \tfrac{1}{2}\log[\det(\underline{V})] + \tfrac{1}{2}\,\underline{e}^T\underline{V}^{-1}\underline{e}. \tag{2.10}$$

If all the entries of $\underline{V}$ are known, then the first two terms of Equation 2.10 are constants, and the MLE method is equivalent to minimizing the final term, or by solving

$$\min_{\underline{p}}\{S(\underline{p}) = \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}[y_i - f(x_i,\underline{p})]V_{ij}^{-1}[y_j - f(x_j,\underline{p})]\}. \tag{2.11}$$

Usually, $\underline{V}$ is not known in advance, but it is often the case that $\underline{V}$ is proportional to a known matrix $\underline{Q}$, and

$$\underline{V} = \rho^2\underline{Q} \tag{2.12}$$

(ρ being unknown). Then Equation 2.10 becomes

$$-\log L(\rho,\underline{p}) = \frac{n}{2}\log 2\pi + n \log \rho + \frac{n}{2}\log[\det(\underline{Q})]$$

$$+ \frac{1}{2\rho^2}\,\underline{e}^T\underline{Q}^{-1}\underline{e}\ . \qquad (2.13)$$

Minimization of -log L with respect to the parameters $\underline{p}$ is equivalent to the minimization problem

$$\min_{\underline{p}}\{S(\underline{p}) = \tfrac{1}{2}\,\underline{e}^T\underline{Q}^{-1}\underline{e}\}. \qquad (2.14)$$

We may compute an estimate $\hat{\rho}$ of the unknown parameter ρ by minimizing -log L with respect to ρ. This yields, from Equation 2.13

$$\frac{\partial L}{\partial \rho} = \frac{n}{\rho} - \frac{1}{\rho^3}\,\underline{e}^T\underline{Q}^{-1}\underline{e} = 0. \qquad (2.15)$$

This yields the estimate

$$\hat{\rho}^2 = \frac{1}{n}\,\underline{e}^T\underline{Q}^{-1}\underline{e}\ . \qquad (2.16)$$

It turns out that the estimate given by Equation 2.16 is biased (Bard, 1974, p. 65). This bias may be approximately eliminated by replacing n in Equation 2.16 by (n-m) (the 'number of degrees of freedom' associated with the estimate) yielding

$$\hat{\rho}^2 = \underline{e}^T\underline{Q}^{-1}\underline{e}\ /(n-m). \qquad (2.17)$$

We further specialize to the Type A and Type B statistical models of Chapter 1, which both satisfy Equation 2.12. For the Type A model $\underline{V} = \sigma^2\underline{I}$ and hence,

$$\min_{\underline{p}}\{S(\underline{p}) = \tfrac{1}{2}\sum_{i=1}^{n}[y_i - f(x_i,\underline{p})]^2\} \tag{2.18}$$

determines the parameters $\hat{\underline{p}}$, and the estimate of σ^2 is given by

$$\hat{\sigma}^2 = \sum_{i=1}^{n}[y_i - f(x_i,\underline{p})]^2/(n-m). \tag{2.19}$$

For the Type B model, $\underline{V} = \sigma^2$ diag $(y_i^t)^2$, which is the diagonal matrix with entries $V_{ii} = (\sigma y_i^t)^2$. We replace y_i^t (the true value of the response function y_i) by $f(x_i,\underline{p})$. Hence Equation 2.14 implies that

$$\min_{\underline{p}}\{S(\underline{p}) = \tfrac{1}{2}\sum_{i=1}^{n}[y_i/f(x_i,\underline{p}) - 1]^2\} \tag{2.20}$$

determines the parameters $\hat{\underline{p}}$, and the estimate of σ^2 is given by

$$\hat{\sigma}^2 = \sum_{i=1}^{n}[y_i/f(x_i,\underline{p}) - 1]^2/(n-m). \tag{2.21}$$

To summarize, *the MLE procedure in the case of normally distributed measurement errors* yields:

Case (i): Covariance matrix of errors known in $\underline{V}$ advance

- use Equation 2.11 to determine $\hat{\underline{p}}$.

Case (ii): $\underline{V} = \rho^2\underline{Q}$

- use Equation 2.14 to determine $\hat{\underline{p}}$
- use Equation 2.17 to estimate $\hat{\rho}^2$

Type A Statistical Model, $\underline{V} = \sigma^2\underline{I}$:

- use Equation 2.18 to determine $\hat{\underline{p}}$
- use Equation 2.19 to estimate $\hat{\sigma}^2$

Type B Statistical Model, $\underline{V} = \sigma^2$ diag $(y_i^t)^2$:

- use Equation 2.20 to determine $\hat{\underline{p}}$
- use Equation 2.21 to estimate $\hat{\sigma}^2$.

One can always apply the MLE procedure if the joint probability density function of the errors is known. Typically, one only needs to know the general form of this function and in some cases, certain of the underlying distributional parameters $\underline{\theta}$ can also be estimated. Under relatively mild conditions on the form of the likelihood function, the MLE method is appropriate in a statistical sense - when the number of observations n is large (Jennrich, 1969; Bard, 1974); and apparently produces acceptable estimates in many situations (Bard, 1974). Better estimates may sometimes be obtained in particular cases using alternative techniques (see, for example, Saleh and Choudhry, 1974; Minder and McMillan, 1977), but its generality and ease of application are powerful arguments for using it in general.

2.2 Bayesian Estimation

Often the scientist has some ideas concerning the likely values of the parameters in his model even before any experiments are performed. It is often the case, for example, that negative values of certain parameters are unacceptable. Thus for Equation 1.8, positive values of β_i are unacceptable if the model describes dissipation of material (f) with time (x).

A similar situation occurs when the scientist has already performed a number of experiments on the system and further experiments are planned.

The information learned about the parameters from the first set of experiments is prior information which can be used in planning the future experiments.

We can think of the prior information concerning the parameters as being characterized in general by a *prior density function* $\phi_0(\underline{p})$. For example, in the case of physically impossible parameter sets $\underline{p}_0$ we must have $\phi_0(\underline{p}_0) = 0$. If we admit the existence of a prior density function $\phi_0(\underline{p})$ then we can combine this with the likelihood function L discussed in Section 2.1 to obtain the *posterior density* function $\phi(\underline{p})$, defined by

$$\phi(\underline{p}) = C\, L(\underline{p})\, \phi_0(\underline{p}) \tag{2.21}$$

where C is a constant, whose value is often immaterial.

According to the arguments of Section 2.1, we should determine $\hat{\underline{p}}$ by maximizing the posterior density function at the observed residuals e_i. This is called the method of maximizing the posterior density function (MPD method) (see, for example, Box and Draper, 1965; Draper and Hunter, 1967; Bard, 1974).

If we have no preference for any particular parameter values over any others, then the prior distribution is uniform, ϕ is proportional to $L(\underline{p})$, and maximizing $\phi(\underline{p})$ is equivalent to maximizing $L(\underline{p})$. If certain parameters $\{p_j\}$ must be non-negative but otherwise arbitrary, then maximizing $\phi(\underline{p})$ is equivalent to the constrained maximization problem

$$\max_{p_j \geq 0} L(\underline{p}). \tag{2.22}$$

A more general case occurs when the parameters are subject to other than non-negativity constraints but otherwise have uniform prior density. This

leads to general problems of the form

$$\max_{\substack{g_k(\underline{p})=0 \\ h_j(\underline{p})\geq 0}} L(\underline{p}). \tag{2.23}$$

A simple example of equality constraints occurs when a subset of $\underline{p}$ represents mole fractions of certain chemicals in some phase. Then for these parameters we must have

$$\sum_k p_k - 1 = 0. \tag{2.24}$$

2.3 Least-Squares Estimation

This method is often used in practice with little or no statistical justification. In this section we consider how least-squares estimation (LSE) arises properly as a special case of the MLE method of Section 2.1.

One of the simplest cases of the MLE method occurs when the errors are independently normally distributed with means of zero and equal variances σ^2 (Type A error behavior of Section 1.4). We have seen in Section 2.1 that this leads to the problem

$$\min_{\underline{p}}\{S(\underline{p}) = \tfrac{1}{2} \sum_{i=1}^{n} [y_i - f(x_i,\underline{p})]^2\}. \tag{2.25}$$

This is the simple least-squares method discussed in elementary statistics texts, sometimes called the method of unweighted least squares (ULS).

A more involved situation occurs when the error variances are unequal.

The error covariance matrix is

$$\underline{V} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & \\ 0 & \sigma_2^2 & & \\ & & \ddots & \\ 0 & \cdots & 0 & \sigma_n^2 \end{pmatrix}. \tag{2.26}$$

Equation 2.11 of the MLE method then becomes

$$\min_{\underline{p}}\{S(\underline{p}) = \tfrac{1}{2} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} [y_i - f(x_i,\underline{p})]^2\}. \tag{2.27}$$

For Type B error behavior (Section 1.4), Equation 2.27 applies with σ_i being replaced by $f(x_i,\underline{p})$. Equation 2.27 is sometimes called the 'method of weighted least-squares' (WLS), since one can view the residuals in the sum as being weighted according to the inverse of their variance σ_i^2.

The most general case occurs when $\underline{V}$ is not diagonal and the elements of its inverse are given by

$$(\underline{V}^{-1})_{ij} = W_{ij}. \tag{2.28}$$

Then, Equation 2.11 of the MLE method becomes

$$\min_{\underline{p}}\{S(\underline{p}) = \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} [y_i - f(x_i,\underline{p})][(y_j - f(x_i,\underline{p})]\}. \tag{2.29}$$

Equation 2.29 is thus a generalization of Equation 2.27, which in turn is a generalization of Equation 2.25.

Thus, the least-squares method is equivalent to the MLE method *when the measurement errors are normally distributed.* When this is not true, the two methods are no longer equivalent. The least-squares method can

still be applied in an *ad hoc* manner to the problem, but the statistical interpretation of the results is then unsatisfactory. However, in many interpolation models, especially those arising in the social sciences, one can often do no better.

Recalling our discussion of Section 1.3, we now see exactly why linear model functions are especially simple in the context of the MLE method and normally distributed measurement errors. In these contexts the models yield a quadratic objective function $S(\underline{p})$ in Equation 2.11.

2.4 Linearizing Transformations

We have emphasized in Sections 1.3 and 2.3 that model functions which are linear in the model parameters are especially convenient when the measurement errors are normally distributed. In view of this, it would appear desirable to employ linear models whenever possible.

Sometimes, the original nonlinear model may be linearized by means of a transformation of the model equation. For example, if we take logarithms of the model

$$y = p_1 \exp(p_2 x) \tag{2.30}$$

we have

$$\log y = \log p_1 + p_2 x \ . \tag{2.31}$$

Defining new variables and parameters by

$$z = \log y$$

$$p_1^* = \log p_1 \tag{2.32}$$

Equation 2.31 becomes

$$z = p_1^* + p_2 x \tag{2.33}$$

which is now a linear model.

Equation 2.33 may now be treated as a linear least-squares problem. However, this may be a totally inappropriate procedure, depending on the statistical properties of the measurement errors accompanying Equation 2.30 (i.e. the statistical model). This has been noted many times in the literature (c.f., Silvert, 1979). For example, if we have

$$y_i = p_1 \exp(p_2 x_i) + e_i \tag{2.34}$$

the corresponding transformed equation is

$$z_i = \log y_i = \log[p_1 \exp(p_2 x_i) + e_i]$$

$$= \log p_1 + p_2 x_i + \log\{1 + \frac{e_i}{[p_1 \exp(p_2 x_i)]}\}. \tag{2.35}$$

The error term in Equation 2.35 thus depends on the (unknown) parameters as well as the value of the independent variable.

If the least-squares method is appropriate for Equation 2.34, it is not appropriate for the transformed equation. However, one can account for the transformed error distribution in the following *approximate* way. Provided the errors are small, one can expand the logarithmic term and use Equation 2.34 to write Equation 2.35 as

$$\log y_i = \log p_1 + p_2 x_i + \frac{e_i}{y_i} . \tag{2.36}$$

If the errors e_i in Equation 2.34 are independent and normally distributed with means 0 and variances σ_i^2, then the errors of Equation 2.36 are approximately independent and normally distributed with means 0 and variances σ_i^2/y_i^2.

Use of the MLE method for Equation 2.34 entails the minimization of the objective function

$$S_1(\underline{p}) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} [y_i - p_1 \exp(p_2 x_i)]^2. \tag{2.37}$$

Use of the MLE method for Equation 2.36 entails the minimization of the objective function

$$S_2(\underline{p}) = \sum_{i=1}^{n} \frac{y_i^2}{\sigma_i^2} (\log y_i - p_1^* - p_2 x_i)^2. \tag{2.38}$$

The weights in the objective function of Equation 2.37 have thus been altered from σ_i^{-2} to $y_i^2 \sigma_i^{-2}$.

In general, if the transformation function $\phi(y)$ produces a linear model, then Equation 1.3 becomes

$$\bar{y}_i = \phi[f(x_i, \underline{p}) + e_i]. \tag{2.39}$$

Equation 2.39 may be approximated by a Taylor expansion about $e_i = 0$ and the replacement of $f(x_i, \underline{p})$ by y_i to yield

$$\bar{y}_i = \phi[f(x_i,p)] + e_i\ \phi'(y_i). \tag{2.40}$$

Then, if the errors e_i are normally distributed with covariance matrix $\underline{V}$, the covariance matrix of the errors resulting from Equation 2.40 is given by

$$\bar{V}_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\partial\phi}{\partial y}\right)\Big|_{y_i} V_{ij} \left(\frac{\partial\phi}{\partial y}\right)\Big|_{y_j} \tag{2.41}$$

where $\Big|_{y_i}$ denotes evaluation at y_i.

We emphasize the fact that Equations 2.40 and 2.41 are *approximations only* and thus must be used with care. This approximate nature does not seem to be universally recognized, however (see, for example, Cvetanovic and Singleton, 1977).

Linearizing transformations *are* actually useful in practice, however, in providing numerical *estimates* of $\hat{\underline{p}}$ to aid the computation in the case of the original nonlinear problem. We shall see a specific example of this in Chapter 7.

CHAPTER 3

NUMERICAL COMPUTATION OF THE PARAMETER ESTIMATES

3.1 INTRODUCTION

We have seen in Chapter 2 that computation of the parameter estimates $\hat{\underline{p}}$ using the MLE method leads to minimization problems of the form

$$\min_{\underline{p}} \{-\log L(\underline{p},\underline{\theta})\} \tag{3.1a}$$

such that

$$h_j(\underline{p}) \geq 0;\ j=1,2,\ldots,n_1 \tag{3.1b}$$

$$g_k(\underline{p}) = 0;\ k=1,2,\ldots,n_2. \tag{3.1c}$$

Equations 3.1b and 3.1c express constraints on the parameters. In typical problems n_1 and n_2 are small (or zero). One case that sometimes arises is the imposition of simple non-negativity constants on the parameters

$$p_j \geq 0 . \tag{3.2}$$

When the likelihood function is nonlinear in $\underline{p}$ and/or the constraints are nonlinear in $\underline{p}$, Equation 3.1 constitutes what is called a *nonlinear programming problem.*

This is an active area of research (Walsh, 1977) and many particular numerical techniques have been applied to parameter estimation problems. It is probably true that no single numerical algorithm has emerged which is superior to all others. Chambers (1977) lists sources from which some computer programs may be obtained.

The numerical solution of Equation 3.1 is often fraught with many

difficulties. One of these is the fact that in practice we can usually only compute *local* solutions. If the objective function and the constraints are convex this guarantees uniqueness, but it is rarely possible to establish this for typical parameter estimation problems. For this reason, it is often advisable to solve the problem numerically using more than one algorithm in order to check for the possibility of nonuniqueness.

Recent texts by Bard (1974), Seinfeld and Lapidus (1974), and Chambers (1977) address themselves to the problem of numerical computation of the parameters. In this chapter, we first wish to outline briefly the basis for the various types of methods currently in use and make some practical suggestions. Second, we will discuss a few very simple algorithms that can be programmed by the scientist himself either on a computer or a hand-held calculator. We will focus here exclusively on the problem of the general minimization of an objective function $S(\underline{p})$ in the absence of constraints. See the above references for discussion of the constrained problem.

All numerical algorithms for solving this problem are iterative in nature. That is, they proceed from some initial estimate of the solution $\underline{p}^0$ and compute new estimates on each iteration by means of

$$\underline{p}^{k+1} = \underline{p}^k + \lambda_k \underline{d}_k \tag{3.3}$$

where λ_k is called a *relaxation parameter* and $\underline{d}_k$ is a *search vector*. Usually λ_k is chosen at each iteration so that the objective function satisfies

$$S(\underline{p}^{k+1}) < S(\underline{p}^k). \tag{3.4}$$

Such an algorithm is said to be *stable*. [Note that this in itself does not guarantee convergence (Broyden, 1967, p. 372).] Equation 3.4 is often achieved by choosing λ_k to minimize $S(\underline{p}^{k+1})$, with $\underline{p}^{k+1}$ defined by Equation 3.3.

Algorithms differ primarily in the way they choose the search vector $\underline{d}_k$. We will briefly discuss three classes of methods. The first are called *search methods*, and they use only *values* of $S(\underline{p})$ on each iteration. The second are called *derivative methods* and use computations of first or second *derivatives* of S with respect to $\underline{p}$ to determine $\underline{d}_k$. The third class is called *differential equation methods*, and do not use Equation 3.3 directly.

3.2 Search Methods

These algorithms make no use of the derivatives of the objective function with respect to the parameters, and use only selected values of $S(\underline{p})$ in the course of the computation. Historically, the earliest of these algorithms is due to Hooke and Jeeves (1961). We will discuss here two of these methods, both of which we recommend when the number of parameters is less than 5. One can be programmed on some hand-held calculators, and the other is a recently developed algorithm that is easily programmed for an electronic computer.

(a) Simplex Method

This method is due to Spendley *et al.* (1962) and Nelder and Mead (1965) and is also discussed by Parkinson and Hutchinson (1972) and by Walsh (1977). It is capable of being programmed on a hand-held calculator and is suitable for problems with a small number of variables (<5). A FORTRAN computer program is given by O'Neill (1971) (see Chambers and

Ertel, 1974, for corrections to this paper). The basic ideas can best be illustrated using two dimensions, and the generalization to higher dimensions is discussed in the above references.

A two-dimensional simplex is a triangle with vertices at the points $\underline{x}_1, \underline{x}_2, \underline{x}_3$. The function values at each vertex are computed. The vertex having the largest function value is rejected and a new vertex is generated by reflecting the rejected point in the opposite side through the centroid of the triangle. The process is repeated with the three new vertices, again rejecting the vertex with the largest function value and determining a new vertex. The simplex is expanded or contracted when necessary, and can also be changed in shape (see references above).

(b) Luus-Jakola Method

This is a recently-developed algorithm (Luus and Jakola, 1973; Jakola and Luus, 1974) which is very easy to program for an electronic computer (a program listing is given in Jakola and Luus, 1974).

It proceeds by using some initial solution estimate $\underline{p}^0$, and uses pre-assigned intervals r^0 and a reduction factor ε (typically $\varepsilon \simeq 0.05$). A large number ($\simeq$2000) of pseudo-random numbers is generated, and these in turn are used to calculate test coordinates $\underline{p}^k$, where

$$p_i^k = p_i^0 + r_i^0 \, y_{ki}, \quad i=1,2,\ldots,m; \; k=1,2,\ldots,\ell \tag{3.5}$$

where m is the number of parameters, $m\ell \simeq 2000$ and y_{ki} are the pseudo-random numbers. The smallest among the resulting function values $S(\underline{p}^k)$ is determined, $\underline{r}^0$ is set to $(1 - \varepsilon)\underline{r}^0$, and the process is repeated until convergence is achieved.

3.3 Derivative Methods

These algorithms use information concerning either the first or second derivatives of $S(\underline{p})$. Since there are m^2 second derivatives and only m first derivatives, much interest has been concerned in recent years with the development of methods that only actually compute first-derivatives, and use this information to construct approximations to the second-derivative matrix. These are called Quasi-Newton methods (Broyden, 1967; Walsh, 1977). Another approach is to use the so-called Gauss-Newton method (Bard, 1974). All these methods are related to the Newton-Raphson method for solving the necessary conditions for the minimization problem, given by

$$\frac{\partial S}{\partial \underline{p}} = \underline{0}. \tag{3.6}$$

We will thus briefly discuss the Newton-Raphson method and show the connection with the Quasi-Newton and Gauss-Newton algorithms.

An algorithm is called a *descent method* if the derivative of S in the direction $\underline{d}_k$ is negative when $\lambda_k = 0$ in Equation 3.3. This ensures that some $\lambda_k > 0$ may be found which will guarantee a decrease of S on each iteration. This derivative is given by

$$\left(\frac{dS}{d\lambda_k}\right)_0 = \frac{d}{d\lambda_k} S(\underline{p}^k + \lambda_k \underline{d}_k)\Big|_{\lambda_k = 0} = \left(\frac{\partial S}{\partial \underline{p}}\right)^T_{\underline{p}^k} \underline{d}_k. \tag{3.7}$$

The Newton-Raphson method expands Equation 3.6 about an estimate $\underline{p}^k$ in a first-order Taylor series. Alternatively, the method may be viewed as that of minimizing the second-order Taylor series approximation to $S(\underline{p})$ near $\underline{p}^k$. This expansion gives

$$S(\underline{p}) \approx S(\underline{p}^k) + \left(\frac{\partial S}{\partial \underline{p}}\right)_0^T \delta\underline{p} + \tfrac{1}{2}\, \delta\underline{p}^T \left(\frac{\partial^2 S}{\partial \underline{p}^2}\right)_0 \delta\underline{p} \tag{3.8}$$

where subscript 0 denotes evaluation at $\underline{p}^k$,

$$\underline{d}_k \equiv \delta\underline{p} = \underline{p} - \underline{p}^k \qquad (cf.\ \text{Eq. } 3.3) \tag{3.9}$$

and

$$\underline{H} \equiv \frac{\partial^2 S}{\partial \underline{p}^2} = \left(\frac{\partial^2 S}{\partial p_i \partial p_j}\right) \tag{3.10}$$

is called the Hessian matrix of S.

Minimization of the right-hand side of Equation 3.8 yields

$$\underline{d}_k = \delta\underline{p} = -\left(\frac{\partial^2 S}{\partial \underline{p}^2}\right)_0^{-1} \left(\frac{\partial S}{\partial \underline{p}}\right)_0 . \tag{3.11}$$

Equation 3.3 with $\underline{d}_k$ given by Equation 3.11 and $\lambda_k = 1$ is called the *unmodified* Newton-Raphson method. The *descent* Newton-Raphson method combines Equation 3.11 with Equation 3.3 to give the iteration

$$\underline{p}^{k+1} = \underline{p}^k - \lambda_k \left(\frac{\partial^2 S}{\partial \underline{p}^2}\right)_0^{-1} \left(\frac{\partial S}{\partial \underline{p}}\right)_0 \tag{3.12}$$

where λ_k is a relaxation parameter.

For the descent Newton-Raphson method, we have, using Equation 3.7,

$$\left(\frac{dS}{d\lambda_k}\right)_0 = -\left(\frac{\partial S}{\partial \underline{p}}\right)_0^T \left(\frac{\partial^2 S}{\partial \underline{p}^2}\right)_0^{-1} \left(\frac{\partial S}{\partial \underline{p}}\right)_0 . \tag{3.13}$$

Thus, provided the Hessian is a positive-definite matrix (which is the case

at the minimum) the descent Newton-Raphson method of Equation 3.12 is a descent algorithm. However, practical difficulties can arise when $\underline{H}_0 = (\partial^2 S/\partial \underline{p}^2)_0$ is indefinite during the course of the computation or when the rows of $\underline{H}_0^{-1}$ are orthogonal to the gradient vector $(\partial S/\partial \underline{p})_0$. Various devices can be used to overcome this problem, one of the simplest being the alteration of Equation 3.12 to

$$\underline{p}^{k+1} = \underline{p}^k - \lambda_k[\omega \underline{I} + \left(\frac{\partial^2 S}{\partial \underline{p}^2}\right)_0]^{-1}\left(\frac{\partial S}{\partial \underline{p}}\right)_0 \tag{3.14}$$

where ω is chosen large enough so that a descent algorithm results.

The unmodified Newton-Raphson method has a desirable 'guaranteed convergence property' in the sense that, if the initial estimate $\underline{p}^0$ is 'close enough' to a solution of Equation 3.6, then the iteration defined by Equation 3.12 with $\lambda_k = 1$ converges to that solution (Jensen and Rowland, 1975). The practical difficulty is that it is usually impossible to test the 'close enough' criterion.

Another practical difficulty with using the (unmodified or descent) method arises from the fact that the Hessian matrix must be recomputed on each iteration and a set of m linear equations solved. These steps can be very time-consuming when m is large, and the Quasi-Newton algorithms attempt to circumvent these difficulties.

In the special case when $S(\underline{p})$ is a sum of squares, as is often the case for the MLE method (see Chapter 2), we have

$$S(\underline{p}) = \tfrac{1}{2} \sum_{k=1}^{n} w_k[y_k - f(x_k, \underline{p})]^2 \tag{3.15}$$

$$\frac{\partial S}{\partial p_i} = - \sum_{k=1}^{n} w_k [y_k - f(x_k, \underline{p})] \left(\frac{\partial f_k}{\partial p_i} \right) \tag{3.16}$$

$$H_{ij} = \frac{\partial^2 S}{\partial p_i \partial p_j} = \sum_{k=1}^{n} w_k \left(\frac{\partial f_k}{\partial p_i} \right) \left(\frac{\partial f_k}{\partial p_j} \right) - \sum_{k=1}^{n} w_k [y_k - f(x_k, \underline{p})] \left(\frac{\partial^2 f_k}{\partial p_i \partial p_j} \right) \tag{3.17}$$

where f_k denotes $f(x_k, \underline{p})$. We thus see that using the Newton-Raphson method entails the calculation of nm^2 second derivatives of the model function in the second term of Equation 3.17.

The Gauss-Newton method ignores the second term of Equation 3.17, and is thus defined by

$$\underline{p}^{k+1} = \underline{p}^{k} - \lambda_k \underline{S}_0^{-1} \left(\frac{\partial S}{\partial \underline{p}} \right)_0 \tag{3.18}$$

where we have defined

$$S_{ij} = \sum_{k=1}^{n} w_k \left(\frac{\partial f_k}{\partial p_i} \right) \left(\frac{\partial f_k}{\partial p_j} \right). \tag{3.19}$$

We call equation 3.18 with $\lambda_k = 1$ the unmodified Gauss-Newton method, and Equation 3.18 the descent GaussNewton method. The latter was apparently first suggested by Hartley (1961).

The derivative of S in the direction defined by Equation 3.18 is given by Equation 3.7 as

$$\left(\frac{dS}{d\lambda_k} \right)_0 = - \left(\frac{\partial S}{\partial \underline{p}} \right)_0^T (\underline{S}_0^{-1})(\underline{S}_0^{-1}) \left(\frac{\partial S}{\partial \underline{p}} \right)_0. \tag{3.20}$$

Thus, provided $\underline{S}_0$ is of full rank everywhere during the computation,

the modified Gauss-Newton method is a descent algorithm. As in the case of Equation 3.14 this can be ensured by modifying Equation 3.18 to

$$\underline{p}^{k+1} = \underline{p}^{k} - \lambda_k(\omega\underline{I} + \underline{S}_0)^{-1} \left(\frac{\partial S}{\partial \underline{p}}\right)_0 . \tag{3.21}$$

Equation 3.21 is often called the 'Marquardt compromise' (Marquardt, 1963) and a computer program based on this algorithm has been available for a number of years from the SHARE Program Library Agency as program #7040 G2 3094. (See Bard, 1974, pp. 323-324, for a list of this and other available nonlinear parameter estimation programs; see also Chambers, 1977.) Inspection of Equation 3.17 reveals that neglect of the second term is generally reasonable if the model fits the data well (i.e. all $[y_k - f(x_k,\underline{p})]$ are small). However, if this term is large, the unmodified Gauss-Newton method no longer possesses the guaranteed convergence property of the Newton-Raphson method (Dennis, 1977). That is, no matter how close the initial estimate $\underline{p}^0$ is to a solution of Equations 3.6, it may be the case that convergence cannot be achieved. Although the descent method may still converge in such cases we may expect convergence to be very slow in practice, especially when the neglected term in Equation 3.17 is large.

Quasi-Newton algorithms use the iteration

$$\underline{p}^{k+1} = \underline{p}^{k} - \lambda_k \underline{B}_k^{-1} \left(\frac{\partial S}{\partial \underline{p}}\right)_k \tag{3.22}$$

where $\underline{B}_k$ is an approximation to the Hessian $\underline{H}_k$. These methods have the property of *quadratic convergence*, which means that they will minimize a strictly quadratic objective function of m variables in m (or less) applications of Equation 3.22. Since objective functions are not strictly quadratic in

practice, the iteration of Equation 3.22 is 'restarted' every m or so iterations.

$\underline{B}_k^{-1}$ is constructed from $\underline{B}_{k-1}^{-1}$ in a computationally efficient manner, thus eliminating the drawback of the Newton-Raphson method that requires the solution of a set of linear equations on each iteration. Quite often, $\underline{B}_0$ is set to the identity matrix, which is readily inverted.

Research is continuing on the development of algorithms related to the Quasi-Newton algorithms. Some recent references are Betts (1976), Davidon (1976), Wolfe (1976), Dennis (1977), Dennis and Moré (1977), and Gill and Murray (1978).

3.4 Differential Equation Methods

Since function minimization is a rather difficult problem and much more is known about the problem of numerical solution of initial value problems involving sets of ordinary differential equations, some recent work has attempted to reformulate minimization problems as sets of ordinary differential equations. Such algorithms are a fairly recent development (Branin, 1972; Villiers and Glasser, 1977; Goh, 1978, 1979), but they may eventually become quite useful for parameter estimation problems. This is due to the fact that there are available many excellent computer packages for solving initial-value ordinary differential equation problems (for example, some of the IMSL routines). The idea is that the user simply 'plugs in' to these packages after reformulating his parameter estimation problem.

One way (Goh, 1979) of formulating the problem of minimizing $S(\underline{p})$ as a differential equation problem is to write

$$\dot{\underline{p}} = -\left[\frac{\partial S}{\partial \underline{p}}\right]$$

$$\underline{p}(0) = \underline{p}_0 \qquad (3.22)$$

where the dot denotes differentiation with respect to time and $\underline{p}_0$ is an initial parameter estimate.

The solution to Equations 3.6 is an equilibrium point $\underline{p}^*$ of the differential Equation 3.22 (i.e. $\dot{\underline{p}} = \underline{0}$ at a solution of Equations 3.6). A simple way of solving Equation 3.6 is thus to employ a computer library routine for solving Equation 3.22. Goh (1979) has discussed general convergence from an arbitrary $\underline{p}_0$ to a solution of Equation 3.6. From the theory of differential equations (Hirsch and Smale, 1974), we know that convergence will occur when $\underline{p}_0$ is 'close enough' to a solution $\underline{p}$ of Equation 3.6, provided the eigenvalues of the Jacobian of the right-hand side of Equation 3.22 have negative real parts. This Jacobian is just the negative of the Hessian matrix of $S(\underline{p})$ and the negativity of the eigenvalues (which are real) requires that the Hessian be positive definite at $\underline{p}^*$. This is just the condition that $\underline{p}^*$ be a local minimum of $S(\underline{p})$.

Goh (1979) has also pointed out a simple way of incorporating non-negativity constraints on the parameters using this approach.

CHAPTER 4

THE PRECISION OF THE PARAMETER ESTIMATES

4.1 The Sampling Distribution of the Estimator

As discussed in Chapter 1, in model fitting we are often interested not only in numerical estimtes $\hat{\underline{p}}$ of the parameters, but also in the precision of these estimates, given the inaccurate experimental data measurements. In the context of linear model functions and normal measurement error probability distribution functions, exact statements can be made concerning the precision of $\hat{\underline{p}}$. In particular, confidence regions and intervals in parameter space can be computed (Draper and Smith, 1966). When the model is nonlinear, these exact results no longer hold. Some discussion of this problem appears in the literature (Beale, 1960; Guttman and Meeter, 1965; Draper and Smith, 1966; Box, 1971; Sorgan and Mikhail, 1971) and we pursue the general aspects of the parameter precision problem in this chapter.

In general, the parameters are determined by means of some functional transformation of the data (Chapter 2) which we denote as

$$\underline{p} = \underline{\phi}(\underline{y}). \tag{4.1}$$

The function $\underline{\phi}$ is called an *estimator* (an example is the MLE method) and $\hat{\underline{p}}$ obtained from Equation 4.1 is called a *point estimate*. The measurement vector $\underline{y}$ is a random variable with some probability density function $f(\underline{y})$. The parameter estimates $\hat{\underline{p}}$ in Equation 4.1 are also random variables. We call their probability distribution the *sampling distribution* of the estimator $\underline{\phi}$. Basically we wish to determine this probability distribution.

The complete computation of this distribution is a formidable computational task, especially when the number of parameters is large (greater than about 3). However, we will discuss in Section 4.6 a simulation technique for accomplishing this. The means and the covariance matrix of the parameters are rather more amenable to approximate calculation, however (see Section 4.2). The covariance matrix can be used to construct approximate confidence regions, as we shall see in Section 4.3. Also in Section 4.3, we shall consider alternate ways of computing approximate confidence regions for $\hat{\underline{p}}$.

If Equation 4.1 is linear in $\underline{y}$ and the measurement errors are normally distributed, then the parameter means and the covariance matrix are linearly related to the corresponding means and covariance matrix of $\underline{y}$ (Draper and Smith, 1966).

In the nonlinear case, the mean of $\hat{\underline{p}}$ is given by the relation

$$\bar{\underline{p}} = E[\hat{\underline{p}}] = E[\underline{\phi}(\underline{y})] = \int \underline{\phi}(\underline{y})\ f(\underline{y})d\underline{y} \tag{4.2}$$

where $E[\cdot]$ denotes the expectation value and the range of the integration is over all allowed values of $\underline{y}$. The distribution of $\underline{y}$ depends on the true values $\underline{y}^t(\underline{p}^*)$ [where $\underline{p}^*$ is the correct (unknown) parameter vector] and on the error distribution. Thus, we can re-write Equation 4.2 as

$$\bar{\underline{p}} = \int \underline{\phi}[\underline{y}^t(\underline{p}^*) + \underline{e}]\chi(\underline{e})d\underline{e}. \tag{4.3}$$

Equation 4.3 illustrates the important practical point that we can usually only calculate an *approximation* to $\bar{\underline{p}}$ (since $\underline{p}^*$ is unknown) by replacing $\underline{p}^*$ by our estimate $\hat{\underline{p}}$ from Equation 4.1. If $\bar{\underline{p}} = \underline{p}^*$, then the estimator $\underline{\phi}$ is

said to be *unbiased*. Most often we desire estimators that are not strongly biased, and the MLE estimator usually satisfies this criterion (Bard, 1974, p. 61) (see also Tukey, 1974).

The covariance matrix of the estimates is given by

$$\underline{V}_p = E[(\hat{\underline{p}}-\bar{\underline{p}})(\hat{\underline{p}}-\bar{\underline{p}})^T] = \int \{\underline{\phi}[\underline{y}^t(\underline{p}^*)+\underline{e}] - \bar{\underline{v}}\}\{\underline{\phi}[\underline{y}^t(\underline{p}^*)+\underline{e}]-\bar{\underline{p}}\}^T \chi(\underline{e})d\underline{e}. \quad (4.4)$$

An estimator is called *efficient* if its variance is as small as possible in some sense (see Bard, 1974, p. 41). The MLE estimator is efficient in the limit of large numbers of observations, and seems to produce acceptable results in many practical situations (see Bard, 1974, p. 61).

Provided we substitute $\hat{\underline{p}}$ for $\underline{p}^*$ and $\bar{\underline{p}}$ in Equation 4.4 and approximate the estimator of Equation 4.1 near $\hat{\underline{p}}$ as a linear function of $\underline{y}$, we can compute approximations to $\underline{V}_p$. We will see how to accomplish this in Section 4.2.

4.2 Approximate Computation of the Covariance Matrix of the Estimates

We can approach the determination of $\underline{V}_p$ directly in an approximate way by considering the normal equations for the parameters and examining their sensitivity with respect to errors in the data. Thus, if we determine the parameter estimates $\hat{\underline{p}}$ from the minimization of some objective function $S(\underline{p},\underline{y})$, we have at $\hat{\underline{p}}$.

$$\frac{\partial S}{\partial \underline{p}} = \underline{0}. \quad (4.5)$$

We now examine the variations $\delta\underline{p}$ of the parameters which solve Equation 4.5

resulting from variations $\underline{\delta y}$ in the data. Differentiating Equation 4.5 we have, to first-order,

$$\left(\frac{\partial^2 S}{\partial \underline{p}^2}\right) \underline{\delta p} + \left(\frac{\partial^2 S}{\partial \underline{p} \partial \underline{y}}\right) \underline{\delta y} = 0 \tag{4.6}$$

where the derivatives are evaluated at the $\hat{\underline{p}}$ which solves Equation 4.5. (Equation 4.6 is exact if S is quadratic in $\underline{p}$ and $\underline{y}$). This gives the variation $\underline{\delta p}$ as

$$\underline{\delta p} \simeq - \underline{H}^{*-1} \left(\frac{\partial^2 S}{\partial \underline{p} \partial \underline{y}}\right) \underline{\delta y} \tag{4.7}$$

where $\underline{H}^*$ is the second-derivative matrix evaluated at the solution to Equation 4.5. The covariance matrix of the parameters is thus approximately given by

$$\underline{V}_p = E[\underline{\delta p} \underline{\delta p}^T] = \underline{H}^{*-1} \left(\frac{\partial^2 S}{\partial \underline{p} \partial \underline{y}}\right) \underline{V} \left(\frac{\partial^2 S}{\partial \underline{p} \partial \underline{y}}\right)^T \underline{H}^{*-1} \tag{4.8}$$

where $\underline{V}$ is the covariance matrix of the data,

$$\underline{V} = E[\underline{\delta y} \underline{\delta y}^T] = E[\underline{e}\underline{e}^T]. \tag{4.9}$$

Equation 4.8 can be applied to any error distribution (statistical model). In the special case of normally distributed errors, we have (*cf.* Eq. 2.11).

$$S(\underline{p}) = \tfrac{1}{2} \underline{e}^T \underline{V}^{-1} \underline{e} \tag{4.10}$$

where

$$e_k = y_k - f(x_k, \underline{p}). \tag{4.11}$$

Then

$$\frac{\partial S}{\partial \underline{y}} = \underline{V}^{-1}\underline{e} \tag{4.12}$$

$$\frac{\partial^2 S}{\partial \underline{p}\partial \underline{y}} = -\underline{B}^T \underline{V}^{-1} \tag{4.13}$$

where $\underline{B}$ is the matrix with elements

$$B_{ij} = \frac{\partial f}{\partial p_j}(x_i, \underline{p}). \tag{4.14}$$

Then Equation 4.8 becomes

$$\underline{V}_p \simeq \underline{H}^{*-1}\underline{B}^T\underline{V}^{-1}\ \underline{B}\ \underline{H}^{*-1}. \tag{4.15}$$

A further simplification takes place if we approximate the Hessian matrix by the first term of Equation 3.17, which yields

$$\underline{H}^* \simeq \underline{B}^T\ \underline{V}^{-1}\ \underline{B}. \tag{4.16}$$

Substituting 4.16 in Equation 4.15 yields

$$\underline{V}_p \simeq \underline{H}^{*-1} = (\underline{B}^T\underline{V}^{-1}\underline{B})^{-1}. \tag{4.17}$$

Finally, the specialization of Equations 4.15 and 4.17 to the Type A and Type B statistical models of Section 1.4 are given in Table 4.1 below.

Table 4.1 Approximate Parameter Covariance Estimates $\underline{V}_p$ for Type A and Type B Statistical Models Resulting from Equations 4.15 and 4.17.

	Equation 4.15	Equation 4.17
Type A (use Eq. 2.19 for $\hat{\sigma}$)	$\underline{H}^{*-1}\underline{B}^T\underline{B}\underline{H}^{*-1}/\hat{\sigma}^2$	$\hat{\sigma}^2(\underline{B}^T\underline{B})^{-1}$
Type B (use Eq. 2.21 for $\hat{\sigma}$)	$\underline{H}^{*-1}\underline{B}^T\text{diag}[f^{-2}(x_k,\hat{\underline{p}})]\underline{B}\underline{H}^{*-1}/\hat{\sigma}^2$	$\hat{\sigma}^2\{\underline{B}^T\text{diag}[f^{-2}(x_k,\hat{\underline{p}})]\underline{B}\}^{-1}$

The entries corresponding to Equation 4.17 are more convenient in practice if the Gauss-Newton method has been employed to compute $\hat{\underline{p}}$ (Section 3.3), in which case $\underline{H}^*$ is usually not available. However, if the Newton-Raphson or Quasi-Newton methods have been employed, $\underline{H}^*$ will be available from the computation.

4.3 Approximate Joint Confidence Regions

If, for any data sample $\underline{y}$, we are able to define a closed, bounded subset of parameter space $S_p(\underline{y})$ in such a way that S contains $\hat{\underline{p}}$ in a fraction γ of all possible data samples; that is,

$$\text{prob}[\hat{\underline{p}} \in S_p(\underline{y})] = \gamma, \tag{4.18}$$

then $S_p(\underline{y})$ is called a γ-joint confidence region for the parameters $\underline{p}$. We can construct approximate confidence regions by using the results of Section 4.2.

First, if the data covariance matrix $\underline{V}$ is known, then $\underline{V}_p$ can be computed from Equation 4.8, 4.15, or 4.17. Then, if we use a normal approximation to the sampling distribution, a $100(1-\gamma)$ percent joint confidence

region for the parameters is defined by (Bard, 1974, p. 188) $\underline{\delta p}$ satisfying

$$\underline{\delta p}^T \underline{V}_p^{-1} \underline{\delta p} \leq \chi_\gamma^2 \tag{4.19}$$

where χ_γ^2 is the upper γ point of the χ^2 distribution with m degrees of freedom. Equation 4.19 defines an ellipsoidal region of parameter space.

Consider now the case when the measurement errors are normally distributed and the data covariance matrix $\underline{V}$ is known only to the extent that

$$\underline{V} = \sigma^2 \underline{Q} \tag{4.20}$$

where σ^2 is an unknown constant and $\underline{Q}$ a known matrix. Note that the Type A and Type B statistical models of Section 1.4 obey Equation 4.20. If the mathematical model equations are linearized about $\hat{\underline{p}}$ then the sampling distribution is approximately normal and a $100(1-\gamma)$ percent joint confidence region is defined by (Bard, 1974, p. 190)

$$\underline{\delta p}^T \underline{B}^T \underline{Q}^{-1} \underline{B}\ \underline{\delta p} \leq m\hat{\sigma}^2 F^\gamma_{m,n-m} \tag{4.21}$$

where $F^\gamma_{m,n-m}$ is the upper γ point of the F distribution with m and n-m degrees of freedom and $\hat{\sigma}^2$ is given by (*cf.* Eq. 2.17)

$$\hat{\sigma}^2 = [\underline{y} - \underline{f}(x,\hat{\underline{p}})]^T \underline{Q}^{-1} [\underline{y} - \underline{f}(x,\hat{\underline{p}})]/(n-m). \tag{4.22}$$

Like Equation 4.19, Equation 4.21 also defines an ellipsoidal region in parameter space. For the special cases of the Type A and Type B statistical models, the approximate joint confidence regions are given by:

Type A: $$\underline{\delta p}^T(\underline{B}^T\underline{B})\underline{\delta p} \leq m\hat{\sigma}^2 F^{\gamma}_{m,n-m} \tag{4.23}$$

Type B: $$\underline{\delta p}^T\{\underline{B}^T \text{ diag}[1/f^2(x_i,\hat{\underline{p}})]\underline{B}\}\underline{\delta p} \leq m\hat{\sigma}^2 F^{\gamma}_{m,n-m}. \tag{4.24}$$

Alternatives to Equations 4.19 and 4.21 are obtained by approximating $S(\underline{p})$ in the neighbourhood of $\hat{\underline{p}}$ by the quadratic Taylor expansion,

$$S(\underline{p}) = S_0 + \left(\frac{\partial S}{\partial \underline{p}}\right)_0 \underline{\delta p} + \tfrac{1}{2}\,\underline{\delta p}^T\,\underline{H^*}\underline{\delta p}. \tag{4.25}$$

Since the first term vanishes at $\hat{\underline{\beta}}$, we have

$$\tfrac{1}{2}\,\underline{\delta p}^T\,\underline{H^*}\underline{\delta p} \simeq S - S_0. \tag{4.26}$$

When the measurement errors are normally distributed, then the approximation of Equation 4.17, when combined with Equation 4.26, yields

$$\underline{\delta p}^T\underline{H^*}\underline{\delta p} = \underline{\delta p}^T\,\underline{V_p}^{-1}\underline{\delta p} \simeq 2(S - S_0). \tag{4.27}$$

Then we may use Equation 4.19 to determine the approximate $100(1-\gamma)$ percent joint confidence region as being defined by the parameter sets satisfying

$$S(\underline{p}) \leq S_0 + \tfrac{1}{2}\,\chi_{\gamma}^2. \tag{4.28}$$

Proceeding in a similar way, we may replace the left-hand side of Equation 4.21 by the approximation $2(S - S_0)$ and hence obtain for both Type A and Type B statistical models, the approximate joint confidence region

$$S(\underline{p}) \leq S(\hat{\underline{p}})\left(1 + \frac{m}{n-m} F^{\gamma}_{m,n-m}\right) \tag{4.29}$$

where $$S(\underline{p}) = \tfrac{1}{2}\,[\underline{y} - \underline{f}(x,\underline{p})]^{T}\underline{Q}^{-1}[\underline{y} - \underline{f}(x,\underline{p})]. \tag{4.30}$$

The confidence regions defined by Equation 4.29 are not in general ellipsoidal; however, they are ellipsoidal and agree with those defined by Equation 4.23 and 4.24 when the linearization of the model about $\hat{\underline{p}}$ is a good approximation.

In practice, Equation 4.29 is only useful when the number of parameters is less than or equal to 3. Only then can the regions be displayed graphically in any reasonable manner (although in higher dimensions two-dimensional cross-sections can still be readily displayed). Although Equations 4.23 and 4.24 also cannot be displayed graphically when the number of parameters is large, the entries of the covariance matrix defining the ellipsoid can be presented, as well as the principal components (see next section). Thus, for three or fewer parameters, we recommend the computation of both types of approximate joint confidence regions. The discrepancy between these will yield a feeling for the reasonableness due to the linearization approximation. For more than three parameters, we recommend only the computation of the covariance matrix and its principal components and eigenvalues.

We present a worked example utilizing the material of this section in Chapter 7.

4.4 Principal Components

The confidence regions defined by Equations 4.19, 4.23, and 4.24 are ellipsoidal in shape in parameter space. A two-dimensional example appears in Figure 6.1. Figures like 6.1 show which parameter combinations are well determined (short axis of ellipse) and which are poorly determined

(large axis). The particular combinations are determined by the eigenvectors of the matrix $\underline{V}_p$ or $\underline{H}^{*-1}$ - see Eq. 4.27).

We may, in general, decompose $\underline{V}_p$ as the product

$$\underline{V}_p = \underline{U}\,\underline{\pi}\,\underline{U}^T \tag{4.31}$$

where $\underline{\pi}$ is the diagonal matrix of eigenvalues π_i of $\underline{V}_p$ and $\underline{U}$ is the unitary matrix whose columns are the normalized eigenvectors of $\underline{V}_p$. Defining a vector of new variables, called the *principal components* of $\underline{V}_p$, by

$$\underline{\rho} = \underline{U}^T\underline{p} \tag{4.32}$$

our estimate of $\underline{\rho}$ is

$$\hat{\underline{\rho}} = \underline{U}^T\,\hat{\underline{p}}\ . \tag{4.33}$$

The covariance matrix of the principal components is thus

$$\underline{V}_\rho = E[\underline{\delta\rho}\,\underline{\delta\rho}^T] = E[\underline{U}^T\underline{\delta p}\,\underline{\delta p}^T\underline{U}] = \underline{U}^T\underline{V}_p\,\underline{U} = \underline{\pi}. \tag{4.34}$$

The principal components are thus uncorrelated with variances π_i.

4.5 Confidence Intervals

In general, we recommend the calculation of the approximate joint confidence regions discussed in Section 4.3. However, if it is desired to determine confidence intervals for individual parameters, these may be obtained from the diagonal entries of $\underline{V}_p$. In the case of the Type A and

Type B statistical models, an approximate 100(1-α) percent confidence interval for p_i is given by

$$p_i = \hat{p}_i \pm t_{\alpha/2} \sqrt{(V_p)_{ii}} \tag{4.35}$$

where $(V_p)_{ii}$ is the i^{th} diagonal element of V_p as given by the second column of Table 4.1, and $t_{\alpha/2}$ is the α/2 critical point of the t distribution with n-m degrees of freedom.

Approximate 100(1-α) percent confidence intervals for the principal components defined by Equation 4.33 are given by

$$\rho_i = \hat{\rho}_i \pm t_{\alpha/2} \sqrt{\pi_i} \tag{4.36}$$

where $t_{\alpha/2}$ is the α/2 critical point of the t distribution with n-m degrees of freedom.

4.6 Exact Computation of the Sampling Distribution

The entire sampling distribution may be generated exactly (apart from the approximation $\underline{p}^* = \hat{\underline{p}}$), in principal, by means of a Monte-Carlo computer simulation technique (Bard, 1974, p. 46; Chambers, 1977, p. 146). This is quite feasible and is recommended when the number of parameters is small (perhaps three or less). It proceeds as follows:

(1) From the given data $\underline{y}_0$ compute $\hat{\underline{p}}$ by the techniques of Chapter 3, using an appropriate estimator $\underline{\phi}(\underline{y}_0)$.

(2) Generate a set of 'pseudo-data' using $\underline{y}_i = \underline{f}(x_k, \hat{\underline{p}}) + \underline{e}_i$, where $\underline{e}_i$ is drawn from the probability distribution assumed for the errors (the statistical model) using a pseudo-random number generator on the computer. Call these data $\underline{y}_i$.

(3) For this set of pseudo-data, compute $\hat{\underline{p}}_i = \underline{\phi}(\underline{y}_i)$.

(4) Repeat steps 2 and 3 as many times as desired.

(5) Obtain the required properties of the sampling distribution by averaging over parameter sets obtained in step 3.

For example, the mean of the sampling distribution is

$$\bar{\underline{p}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\underline{p}}_i . \tag{4.37}$$

The covariance matrix is

$$\underline{V}_p = \frac{1}{N-1} \sum_{i=1}^{N} (\hat{\underline{p}}_i - \bar{\underline{p}}_i)(\hat{\underline{p}}_i - \bar{\underline{p}})^T . \tag{4.38}$$

In both the above equations, $\hat{\underline{p}}_i$ is a parameter set obtained from step 3 and N is the number of times steps 2 and 3 have been repeated. It is readily seen that Equations 4.37 and 4.38 amount to Monte-Carlo computations of Equations 4.3 and 4.4 respectively.

This Monte Carlo procedure has apparently not been used extensively in the literature up to now, but it promises to be an extremely useful tool in evaluating sampling distributions and testing the approximations of Section 4.3, as well as in examining the robustness of estimators.

CHAPTER 5

FURTHER IMPLICATIONS OF MODEL FITTING

5.1 Examination of the Residuals

After the parameters have been estimated using any of the methods discussed in Chapter 3, the model produces a set of predicted responses y_k^p at the observed x_k values given by

$$y_k^p = f(x_k, \hat{\underline{p}}). \tag{5.1}$$

The residuals r_k at x_k are the differences between y_k^p and the actually observed response values, y_k,

$$r_k = y_k - f(x_k, \hat{\underline{p}}). \tag{5.2}$$

If the model (and parameters) are correct, then the r_k can be thought of as the observation errors e_k (assuming $\hat{\underline{p}} = \underline{p}^*$, the 'true' parameter values). Since in computing $\hat{\underline{p}}$ we have made certain assumptions concerning the errors (the statistical model - see Sections 1.4 and 2.1), the observed r_k should be examined to ascertain that they conform to these assumptions. In this section we briefly discuss ways of doing this. Chapter 3 of Draper and Smith (1966) and Endrenyi's (1976) article should be consulted for further details.

As discussed in Section 1.4, the commonest types of error assumptions are that the errors are independently normally distributed with zero mean and constant variance σ^2 (Type A) or with variance $\sigma^2(y_i^t)^2$ (Type B).

One of the most useful ways to examine the error assumptions is to plot the residuals against y_k and check the plot visually (Draper and Smith, 1966; Endrenyi, 1976). Four typical situations are shown in Figures 5.1 to 5.4.

In Figure 5.1, the residuals appear to be distributed in a random manner above and below zero, and lie within a 'horizontal band.' In this case, we have no reason to reject either the model or the assumed error behavior.

In Figure 5.2, Type A error behavior has been assumed and the residuals lie in a widening wedge-shaped zone. This indicates that the error behavior is really Type B, contrary to the assumption. A similar situation occurs if Type B error behavior has been assumed and Type A error behavior actually exists. In this case the residuals lie in a narrowing wedge-shaped zone, like that illustrated in Figure 5.3.

If the residuals show systematic trends (Fig. 5.4) then we have reason to believe that the model itself is inadequate. However, systematic trends like those of Figure 5.4 can also be due to isolated poorly-measured data points (outliers). There are many papers in the statistical literature, especially in the case of linear models, which are devoted to discovering the presence of outliers (for a recent example, see Andrews and Pregibon, 1978).

If we have an independent estimate $\hat{\sigma}^2$ of σ^2, then we can make a statistical test of the significance of the difference between the observed variance of the residuals, given by

$$s^2 = \begin{cases} \sum_{i=1}^{n} r_i^2/(n-m) & \text{(Type A error)} \\ [\sum_{i=1}^{n} r_i^2/f^2(x_i,\underline{\hat{p}})]/(n-m) & \text{(Type B error)} \end{cases} \qquad (5.3)$$

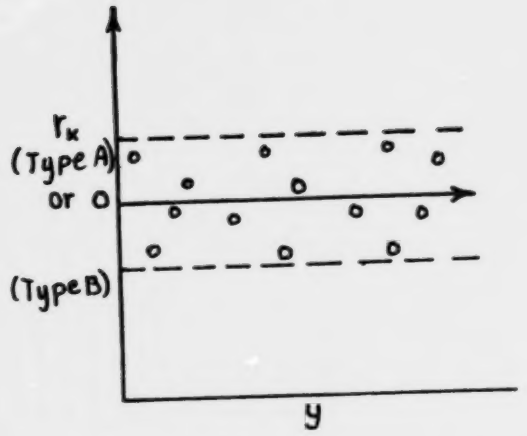

Fig. 5.1 Residuals plotted against dependent variable. Type A or Type B statistical model correctly assumed.

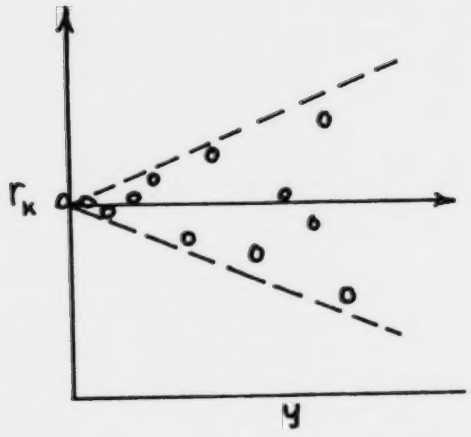

Fig. 5.2 Residual plot. Type A statistical model assumed and Type B actually present.

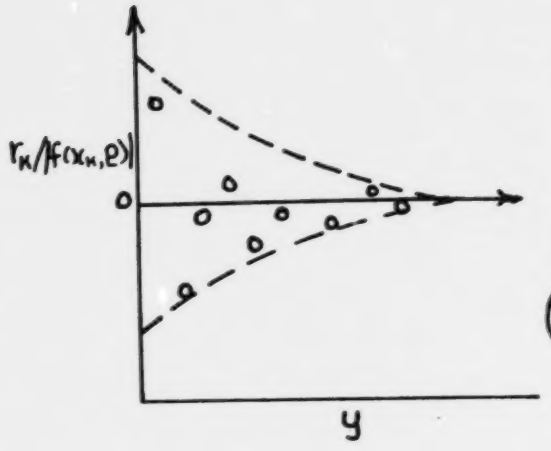

Fig. 5.3 Residual plot. Type B statistical model assumed and Type A actually present.

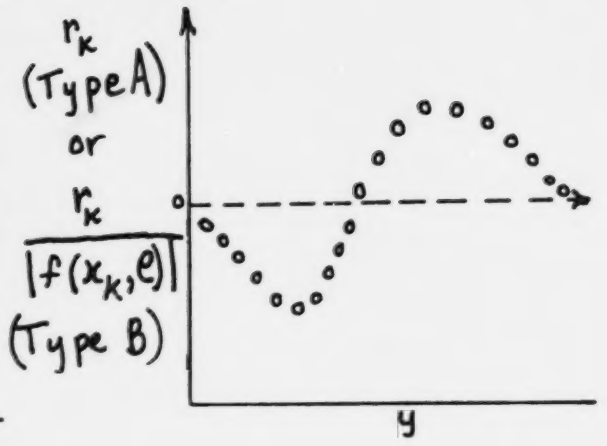

Fig. 5.4 Residual plot. Residuals show systematic trends, indicating possible inadequacy of the mathematical model.

and the estimate of σ^2, by comparing the ratio $s^2/\hat{\sigma}^2$ with the cutoff points of the $F_{n-m,\hat{n}}$ distribution ($\hat{n}$ being the number of degrees of freedom associated with the estimate $\hat{\sigma}^2$).

We usually assume that the mean of the error of the probability density function is zero. Sometimes this assumption may also be tested. However, if the model function contains a constant term, then this guarantees that the mean residual vanishes. For example, if the model is

$$y = a_1 + f(x,\underline{p}) \tag{5.4}$$

and if Type A error behavior is assumed, the normal equation that results from setting the derivative of the sum of squares of the residuals with respect to a_1 equal to zero is

$$\frac{\partial S}{\partial a_1} = -2 \sum_{i=1}^{n} [y_i - a_1 - f(x_i,\underline{p})] = -2 \sum_{i=1}^{n} r_i = 0. \tag{5.5}$$

5.2 The Precision of Model Predictions

We consider the model equation

$$y = f(x,\hat{\underline{p}}) \tag{5.6}$$

and wish to estimate the values of the measured responses at some (hitherto unmeasured) values of x, denoted by the vector $\underline{x}$. There are three sources of inaccuracy in the use of Equation 5.6. These are due to possible experimental inaccuracies in the setting of $\underline{x}$, inaccuracies in the values of $\hat{\underline{p}}$, and measurement errors in y_k the response vector $\underline{y}$.

If the three errors are $\underline{\delta x}$, $\underline{\delta \hat{p}}$, and $\delta \underline{y}$, then the observed values

of $\underline{y}$ will be given by

$$\underline{y} = \underline{f}(\underline{x} + \delta\underline{x}, \underline{p}^* + \delta\hat{\underline{p}}) + \delta\underline{y}. \tag{5.7}$$

The covariance matrix of the observation errors can be estimated by expanding the model function in Equation 5.7 in a first-order Taylor series to yield errors $\underline{e}$ given by

$$\underline{e} = \underline{y} - \underline{y}^0 = \left(\frac{\partial \underline{f}}{\partial \underline{x}}\right)^T \delta\underline{x}_k + \left(\frac{\partial \underline{f}^T}{\partial \hat{\underline{p}}}\right)^T \delta\hat{\underline{p}} + \delta\underline{y} \tag{5.8}$$

where $\underline{y}^0$ is the 'correct' value $\underline{f}(\underline{x},\underline{p}^*)$. Assuming $\delta\underline{x}_k$, $\delta\hat{\underline{p}}$, and $\delta\underline{y}_k$ are independent, and replacing $\underline{p}^*$ by $\hat{\underline{p}}$ in the evaluation, the covariance matrix is given by

$$E[\underline{e}\underline{e}^T] = \left(\frac{\partial \underline{f}^T}{\partial \underline{x}}\right)^T \Bigg|_x \underline{V}_x \left(\frac{\partial \underline{f}^T}{\partial \underline{x}}\right)\Bigg|_x + \left(\frac{\partial \underline{f}^T}{\partial \underline{p}}\right)^T \underline{V}_p \left(\frac{\partial \underline{f}^T}{\partial \underline{p}}\right) + \underline{V}_y \tag{5.9}$$

where $\underline{V}_x$, $\underline{V}_p$, and $\underline{V}_y$ are, respectively, the covariance matrices of $\delta\underline{x}$, $\delta\hat{\underline{p}}$, and $\delta\underline{y}$. If $\underline{x}$ is known precisely, the first term in Equation 5.9 may be omitted. $\underline{V}_p$ in the second-term is obtained using the methods of Chapter 4. $\underline{V}_y$ is either known from the measurement error characteristics, or is estimated from the residuals. The covariance matrix of the prediction errors for the true response from Equation 5.6 is obtained by omitting $\underline{V}_y$ from Equation 5.9. We show an example of the use of Equation 5.9 in Chapter 7.

5.3 Experimental Design

Consider again the model function

$$y = f(x,\hat{\underline{p}}) \tag{5.10}$$

The estimates $\hat{\underline{p}}$ arise from analyzing some set of observational data $\{(x_i,y_i)|$ $i=1,2,\ldots,n\}$. We saw in Chapter 4 how to estimate the uncertainties in the parameters by the computation of confidence regions. To do this, we essentially considered the sensitivity of our estimator to the measured $\underline{y}$ data, by considering Equation 4.5. In fact, the parameter values depend not only on the measurement errors, but also on the values $\{x_i|i=1,2,\ldots,n\}$, in the observation set. The phrase 'design of the experiment' means roughly 'determining the set of x_i values at which the experiments are to be conducted'. Some references to experimental design for nonlinear models are Box and Lucas, 1959; Heineken *et al.*, 1967; Atkinson and Hunter, 1968; Box, 1968, 1970; Hill *et al.*, 1968; Box and Draper, 1972.

Thus, we write our estimator as

$$\underline{p} = \underline{\phi}(x_k,y_k) \tag{5.11}$$

and Equation 5.10 is written as

$$y = f[\underline{x},\underline{\phi}(x_k,y_k)]. \tag{5.12}$$

In many biological situations, the values $\{x_i\}$ at which measurements are taken may be beyond the control of the investigator. However, it may occasionally be possible in some situations to influence the choice of $\{x_i\}$ values used in future experiments. For example, suppose we have the model

$$y = p_1e^{p_2x} \tag{5.13}$$

and have determined tentative values of $\hat{p}_1$ and $\hat{p}_2$ from some preliminary

observation set $\{(x_i,y_i)|i=1,2,\ldots,n\}$. These parameter estimates give rise to confidence intervals as discussed in Chapter 4 and thence to uncertainties in the prediction of y values from Equation 5.13 as discussed in Section 5.2.

Suppose further that additional experiments are to be performed at some set of x_i values $\{x_i|i=n+1,\ldots,n+k\}$. How should we choose those x_i values? i.e., how should we design the experiments? The answer to this question depends on the reason why we are constructing the model in the first place. As discussed in Chapter 1, we may be interested in predicting values of y from Equation 5.10 (prediction), or in obtaining reliable estimates $\hat{\underline{p}}$ of the parameters which can be used in related experimental situations (insight).

If the intent is to minimize some set of prediction errors, then we essentially wish to minimize some function of the covariance matrix of the prediction errors related to Equation 5.9. If the intent is to minimize parameter uncertainties, then we essentially wish to minimize some function of the covariance matrix $\underline{V}_p$ of the parameter estimates discussed in Chapter 4.

As another possibility we may wish to minimize the variance or covariance of some arbitrary function of the properties of the model Equation 5.10. For example suppose we are interested in the maximum value of the response over all x, i.e.

$$y_{max} = \max_x f(x,\underline{p}) \tag{5.14}$$

(assuming such exists). Then y_{max}, which occurs at some x^*, is determined from (assuming f is suitably differentiable)

$$\frac{\partial f}{\partial x}(x,\underline{p}) = \underline{0}. \tag{5.15}$$

Linearization of Equation 5.15 about the solution x* yields

$$\left(\frac{\partial^2 f}{\partial x^2}\right) \delta x + \left(\frac{\partial^2 f}{\partial x \partial \underline{p}}\right) \underline{\delta p} = 0. \tag{5.16}$$

Approximations similar to those applied to Equation 5.7 yield

$$\delta y_{max} \simeq \left(\frac{\partial f}{\partial x}\right) \delta x + \left(\frac{\partial f}{\partial \underline{p}}\right)^T \delta \hat{\underline{p}}$$

$$\simeq \left(\frac{\partial f}{\partial \underline{p}}\right)^T \delta \hat{\underline{p}} \qquad \text{(from Equation 5.15)} \tag{5.17}$$

and hence we have

$$E[\delta y_{max}{}^2] = \underline{\delta \hat{p}}^T \left(\frac{\partial f}{\partial \underline{p}}\right)\left(\frac{\partial f}{\partial \underline{p}}\right)^T \delta \hat{\underline{p}}. \tag{5.18}$$

We would then design our experiments to minimize Equation 5.18.

To summarize, we first identify those quantities whose precision we are interested in - predictions y_k at some x_k values, particular parameter estimate(s), properties predicted by the model function, etc. Then we minimize an appropriate function of the covariance matrix of the desired quantities.

An appropriate 'function of the covariance matrix' that is mentioned above is in general (Bard, 1974, Ch. 10) the logarithm of its determinant. This criterion can be based on information-theoretic arguments, but we accept it here as heuristically reasonable. For normally distributed quantities, $\det^{\frac{1}{2}}(\underline{V})$ is proportional to the volume of a joint confidence

region (Bard, 1974, Section 7.10). Thus, we can approximately minimize the volume of the confidence region for the prediction of the desired quantities by minimizing det $\underline{V}$.

CHAPTER 6

SOME SPECIAL SITUATIONS

6.1 Separable Least-Squares Problems

If the model function is linear with respect to some parameters and nonlinear with respect to others, the resulting least-squares parameter estimation problem is called a *separable least-squares problem.* For example, suppose our model function is

$$f(x,\underline{p}) = p_1 + p_2 e^{p_3 x} \quad . \tag{6.1}$$

This model is linear in p_1 and p_2 but nonlinear in p_3. For model functions of this type, the numerical computations of Chapter 3 can be considerably simplified. Such problems have been discussed extensively in the recent literature (Lawton and Sylvestre, 1971; Barham and Drane, 1972; Guttman *et al.*, 1973; Kaufman *et al.*, 1978).

The least-squares estimation of $\underline{p}$ results from minimizing

$$S(\underline{p}) = \sum_{i=1}^{n} w_i (y_i - p_1 - p_2 e^{p_3 x_i})^2. \tag{6.2}$$

Since the order in which parameters are considered in the minimization is clearly immaterial, we can regard p_3 as the single variable which we must consider in the minimization. For each fixed p_3 value, p_1 and p_2 are determined by linear regression. Thus, we may write

$$S(p_3) = \sum_{i=1}^{n} w_i [y_i - p_1(p_3) - p_2(p_3) e^{p_3 x_i}]^2. \tag{6.3}$$

One way to view Equation 6.3 is as a result of a 'principal of optimality' similar to that used in dynamic programming (Dreyfus and Law, 1977). Alernatively, one may simply recognize that the normal equations resulting from the minimization problem may be partially 'decoupled.' Thus, for Equation 6.2, the normal equations are

$$\begin{pmatrix} \sum_{i=1}^{n} w_i & \sum_{i=1}^{n} w_i e^{p_3 x_i} \\ \sum_{i=1}^{n} w_i e^{p_3 x_i} & \sum_{i=1}^{n} w_i e^{2p_3 x_i} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} w_i y_i \\ \sum_{i=1}^{n} w_i y_i e^{p_3 x_i} \end{pmatrix} \qquad .4)$$

$$p_1 p_2 \sum_{i=1}^{n} w_i x_i e^{p_3 x_i} + p_2^2 \sum_{i=1}^{n} w_i e^{2p_3 x_i} = \sum_{i=1}^{n} w_i y_i e^{p_3 x_i}.$$

The first two members of Equation 6.4 may be solved for p_1 and p_2 in terms of p_3 and the result substituted in the last equation.

Basically, we merely observe that for separable least-squares problems, the dimensionality of the minimization problem which we must solve [i.e. the number of variables with respect to which we must minimize $S(\underline{p})$] is reduced from m by the number of parameters which appear linearly in the model function. This can considerably reduce the labour (either by hand or when using an electronic computer) in effecting a solution using the methods of Chapter 3. The worked example considered in Chapter 7 is a separable least-squares problem.

6.2 Differential Equation and Other Implicit Models

In many cases, the model equation is given not explicitly in the form of Equation 1.1, but in implicit form. A very common example of this is the case where a vector-valued response function $\underline{y}$ is the solution of a set of ordinary differential equations

$$\dot{y}_k \equiv \frac{dy_k}{dt} = f_k(t,\underline{p},\underline{y});\ k=1,2,\ldots,r \tag{6.5}$$

$$y_k(0) = y_k^*.$$

A simple specific example of Equation 6.5 is the Lotka-Volterra system

$$\dot{y}_1 = k_{11}y_1 + k_{12}y_1y_2 \tag{6.6}$$

$$\dot{y}_2 = k_{21}y_2 + k_{21}y_1y_2.$$

We assume here that we cannot solve Equation 6.5 in closed form since, if we could, we would simply have the usual explicit form of model. In general, we must obtain the predicted model responses y by a numerical integration of Equation 6.5. There are many efficient computer packages available for this purpose (for example, several of the IMSL routines).

We suppose we have a set of observations $\{y_i | i=1,2,\ldots,n\}$ corresponding to a set of times t_i. We wish to estimate the parameters in Equation 6.5 by the MLE (or any other) procedure. As we saw in Chapter 3, many algorithms for computing parameter estimates require derivatives of the observations

with respect to the parameters. The purpose of this section is to outline how these can be efficiently calculated.

We can differentiate Equation 6.5 with respect to parameter p_j to obtain

$$\frac{\partial}{\partial p_j}(\dot{y}_k) = \frac{\partial f_k}{\partial p_j} + \sum_{\ell=1}^{r} \left(\frac{\partial f_k}{\partial y_\ell}\right)\left(\frac{\partial y}{\partial p_j}\right) \quad ; \; i=1,2,\ldots r. \tag{6.7}$$

We then rewrite Equation 6.7 as the matrix set of differential equations

$$\frac{d}{dt}\left(\frac{\partial y_k}{\partial p_j}\right) = \frac{\partial f_k}{\partial k_j} + \sum_{\ell=1}^{r} \left(\frac{\partial f_k}{\partial y_\ell}\right)\left(\frac{\partial y}{\partial p_j}\right) \quad ; \quad \begin{matrix} k=1,2,\ldots,r \\ j=1,2,\ldots,m \; . \end{matrix} \tag{6.8}$$

Equations 6.8 are seomtimes called the 'sensitivity equations' for Equation 6.5 (Dickinson and Gelinas,]976; Cukier *et al.*, 1978). They describe the rates of change of a solution to Equation 6.5 with respect to the parameters. Equations 6.8 are in practice integrated simultaneously with Equation 6.5 to determine $y_k(t_i)$ and $(\partial y_k/\partial p_j)$ at t_i. The initial conditions for the quantities in Equation 6.8 are usually zero.

Sometimes one or more of the initial conditions y_j^* in Equation 6.5 are also treated as unknown parameters to be determined. For such parameters, the analogue of Equation 6.8 is

$$\frac{d}{dt}\left(\frac{\partial y_k}{\partial y_j^*}\right) = \sum_{\ell=1}^{r} \left(\frac{\partial f_k}{\partial y_\ell}\right)\left(\frac{\partial y_\ell}{\partial y_j^*}\right) \quad ; \quad k=1,2,\ldots,r. \tag{6.9}$$

The initial conditions for Equation 6.9 are then

$$\left(\frac{\partial y_k}{\partial y_j^*}\right)(t=0) = \begin{cases} 1 & k=j \\ 0 & k\neq j \end{cases} \; . \tag{6.10}$$

For differential equation models, numerical integration of Equation 6.5 often requires great care, especially when the number of equations is large. The phenomenon called 'stiffness' often arises and appropriate numerical methods must be implemented to deal with it. Roughly speaking, a set of differential equations is stiff if the different equations proceed on vastly different time scales. For these systems, time steps in the numerical integration may be required to be very small for the 'fast' equations since the derivatives are so large; whereas for the 'slow' equations, the time steps could be relatively very large. Much computer time can be wasted if the small time steps are used for every equation. Stiff equations typically arise in the context of differential equations describing chemical kinetics, and the term itself was coined by two chemists (Curtiss and Hirschfelder, 1952). There are good numerical algorithms available to handle stiff sets of differential equations, originating from the pioneering work of Gear (1968, 1971). Recent discussions of the problem of stiffness have been given by Bush and Dyer (1976), Warner (1971), and Noyes (1977).

For other implicit models, we can proceed in a way similar to that described above to determine derivatives of the predicted observations with respect to the parameters. For example, consider the implicit model

$$f_k(\underline{y},\underline{x},\underline{p}) = 0;\ k=1,2,\ldots,r \tag{6.11}$$

(we assume $\underline{y}$ is an r dimensional vector). Implicitly differentiating with respect to p_j yields

$$\sum_{\ell=1}^{r} \left(\frac{\partial f_k}{\partial y_\ell}\right)\left(\frac{\partial y_\ell}{\partial p_j}\right) + \left(\frac{\partial f}{\partial p_j}\right) = 0. \tag{6.12}$$

Provided the Jacobian matrix $\partial f_k/\partial y_\ell$ is nonsingular at $\underline{x}_k$, then Equation 6.12 is a set of linear equations yielding $\partial y_\ell/\partial p_j$ at $\underline{x}_k$. Equation 6.11 is in general a set of nonlinear equations which must be solved for $y(x_k)$.

An alternative to numerically solving Equation 6.11 for y_k is to use the model equations as constraints in the minimization of the likelihood function. For example, in the simple case when r=1 and we have the implicit model

$$f(y,\underline{x},\underline{p}) = 0 \tag{6.13}$$

and the Type A statistical model holds, we can introduce new variables y_i^* and solve the constrained minimization problem

$$\min \sum_{i=1}^{n} (y_i - y_i^*)^2 \tag{6.14}$$

$$\text{such that} \qquad f(y_i^*,\underline{x}_i,\underline{p}) = 0;\ i=1,2,\ldots,n.$$

See Britt and Luecke (1973) and Bard (1974) for further discussion of this approach.

6.3 Linear Compartmental Models

A linear compartmental model is a model of the form

$$\underline{y} = \sum_{j=0}^{J} \underline{c}_j e^{-k_j t} \ . \tag{6.15}$$

Linear compartmental models can be obtained from a physical picture consisting of 'boxes' or 'compartments' with flows into and out of each box

which are linear in $\underline{y}$. A simple one-compartment linear model is described by the differential equation

$$\frac{dy}{dt} = c_1 - c_2 y. \tag{6.16}$$

The general solution of Equation 6.16 is of the form given by Equation 6.15,

$$y = c_1 + (y_0 - c_1)e^{-c_2 t}. \tag{6.17}$$

The general linear compartmental model arises from a linear system of ordinary differential equations of the form

$$\frac{dy_i}{dt} = \sum_{k=1}^{r} a_{ik} y_k + b_i;\ i=1,2,\ldots,r. \tag{6.18}$$

The usual objective is to find values for the constants a_{ik} and b_i. It can be shown (see any elementary text on ordinary differential equations) that provided all the eigenvalues of the matrix $\underline{A}$ are real and distinct, each solution y_i of Equation 6.18 can be written in the form of Equation 6.15, where the c_i are related to the a_{ij} and b_k, and the k_j are the eigenvalues of A.

We discuss linear compartmental models here for two reasons. First, they are ubiquitious in biology. Examples are radioactive tracers (Sheppard, 1962), pharmacokinetics (Metzler, 1971, 1974) first-order biochemical kinetics models (Schwartz and Gelb, 1978) and the uptake of chemicals by marine organisms (Branson *et al.*, 1975). The second reason we discuss them here is to emphasize the fact that they have several very unsatisfactory

features, which are not always appreciated by many scientists using such models. The unsatisfactory features are connected with the fact that they often lead to in numerically extremely ill-conditioned parameter estimation problems and the resulting parameter estimates can have very large confidence regions. These numerical difficulties increase with the number of terms used in Equation 6.15, and begin when $J \geq 2$.

In fact, in a recent text on numerical methods, Acton (1977) states concerning persons bringing such problems to the computer laboratory that "the sooner the hopeful innocent can be sent away from the computer room, the better off everyone will be." Our view is not quite as extreme, but we can sympathize with Acton. Part of our conviction in this regard results from the observation that many scientists, on observing a curve $x(t)$ that decreases smoothly with time, will immediately attempt to fit it to Equation 6.15 without thinking about the underlying processes affecting $x(t)$. Just as we emphasized in Chapter 1 that linear models are often used for convenience when in fact they may not be valid, so we emphasize here that Equation 6.15 is also often used for the wrong reasons.

The above having been said, however, compartmental models *are* valid on occasion (especially when the number of exponential terms is small) and they should not be rejected out of hand. If the only objective is to construct a convenient representation of the observed data (i.e. curve-fitting - see Chapter 1) then Equation 6.15 can be very useful. The difficulties arise when one wishes to obtain precise estimates of the parameters (i.e. model-fitting - see Chapter 1, with which this manual is primarily concerned). We will see presently the reason for the difficulties.

Parameter estimation for compartmental models has been discussed in the iterature by Lipton and McGilchrist (1963); Myhill *et al.* (1965); Beauchamp and Cornell (1968, 1969); Cannon and Filmer (1968); McGilchrist (1968);

Parsons (1968); Hildebrand (1974); Swartz and Bremermann (1975); Jennrich and Bright (1976); and Wiscombe and Evans (1977). The final paper is especially recommended, since the method described there apparently overcomes the problem of numerical ill-conditioning (but the large confidence regions associated with the parameters of Equation 6.15 are an inherent property of such models).

The reasons for the ill-conditioning of parameter estimates obtained using Equation 6.15 are probably best illustrated by considering one of the oldest ways used to compute them. This is Prony's method (Hildebrand, 1974; section 9.4) apparently first proposed in 1795. This approach, as well as many others (see previously cited references), assumes that data are available at equally spaced values of t given by

$$t_i = ih;\ i=1,2,\ldots,n. \tag{6.19}$$

For a single response variable Equation 6.15 may then be written as

$$y(t_i) \equiv y_i = \sum_{j=0}^{J} c_j \theta_j{}^i. \tag{6.20}$$

where

$$\theta_j = e^{-k_j h}. \tag{6.21}$$

The theory of linear *difference equations* (Boole, 1960) states that if $n = J+1$, y_i can be made to satisfy exactly a certain homogeneous linear difference equation with θ_j being the zeros of its characteristic polynomial. Then the y_i satisfy the difference equation

$$\sum_{\ell=0} \alpha_\ell\, y_{i+\ell} = 0;\ i=1,2,\ldots,J+1 \tag{6.22}$$

with characteristic polynomial

$$\sum_{\ell=0}^{J} \alpha_\ell \, \theta^\ell = 0. \tag{6.23}$$

The objective of Prony's method is to find $\{\alpha_\ell\}$ and hence $\{\theta_j\}$. It is convenient in Equation 6.22 to let $\alpha_0 = 1$ and hence, provided the resulting coefficient matrix is nonsingular, Equation 6.22 can be solved as a set of linear equations to yield $\{\alpha_1,\ldots,\alpha_J\}$. The solution of the polynomial equation given by Equation 6.23 then yields $\{\theta_j | j=0,\ldots,J\}$.

In the general case when $n > J+1$, we simply find the *least-squares* solution of the set of equations resulting from Equation 6.22. That is, Equation 6.22 is treated as a linear regression problem for the $\{\alpha_\ell\}$. Finally, the $\{c_j\}$ in Equation 6.20 are found by solving the resulting linear regression problem using $\{\theta_j\}$ obtained above.

Prony's method does not find the MLE or least-squares estimates of $\underline{c}$ and $\underline{\theta}$, but is a heuristic way of obtaining 'reasonable' values for them. Inspection of the approach reveals that Prony's method actually finds a set of weights $\{\alpha_k\}$ in Equation 6.22 to minimize

$$\sum_{i=1}^{n} \left(\sum_{\ell=0}^{J} \alpha_\ell y_{i+\ell} \right)^2 = \sum_{i=1}^{n} \left(\sum_{\ell=0}^{J} \alpha_\ell \, \varepsilon_{i+\ell} \right)^2 \tag{6.24}$$

where $\{\varepsilon_i\}$ are the observation errors, yielding a 'best-fitting characteristic polynomial' to the underlying difference equation.

Numerical instability arises for two reasons. First, the linear normal equations resulting from both the determination of $\{\alpha_k\}$ in Equation 6.22 and in the determination of the final $\{c_j\}$ in Equation 6.20 can be

extremely ill-conditioned and great care must be exercised in obtaining solutions. Such instability was mentioned in Chapter 1 for linear least-squares problems in general, and it arises frequently here, especially when J is large. The actual *numerical* difficulties may be overcome (Lawson and Hanson, 1974; Wiscombe and Evans, 1977). However, this ill-conditioning means that small changes in the normal equations (in our case, small errors in $\{y_i\}$ can yield large changes in the solutions to these normal equations (the $\{\alpha_\ell\}$ values). Thus, the $\{\alpha_\ell\}$ often are associated with 'large' confidence regions.

The second reason that Prony's method is often numerically unstable is a reflection of a basic property of Equation 6.23. We have seen that the decay constants $\{k_j\}$ are obtained from $\{\theta_j\}$ in Equation 6.21, which in turn are the zeros of the polynomial Equation 6.23. The coefficients of this polynomial have some uncertainty due to the observational errors, and thus $\{\alpha_\ell\}$ lie within some confidence region (which may be large if the normal equations are ill-conditioned). This imprecision of $\{\alpha_\ell\}$ in turn results in imprecise zeros of the polynomial (i.e. affects *their* confidence regions). It is well known that the dependence of the zeros of polynomials on their coefficients is a classic case of ill-conditioning in numerical analysis (Ralston, 1965). That is, small changes in the coefficients of a polynomial can lead to large changes in its zeros. Hildebrand (1974) on pages 459 to 461 discusses an interesting simple two-exponential example in which very slight changes in the observation errors lead to modest changes in $\{\alpha_\ell\}$ due to the ill-conditioning of the normal equations, which in turn leads to marked changes in $\{\theta_j\}$.

We emphasize finally, however, that even though the *parameters* in Equation 6.20 are often ill-determined (large confidence regions) the

resulting different sets of parameter values can lead to almost equally good fits of the data.

For example (Hildebrand, 1974) the functions

$$f_1(x) = c_{10} + c_{11}e^{-0.96x} + c_{12}e^{-2.18x} \tag{6.25}$$

$$f_2(x) = c_{20} + e^{-1.29x}(c_{21} \cos 0.515x + c_{22} \sin 0.515x) \tag{6.26}$$

lead to equally good representations of the function

$$f(x) = 2.32 - 1.08e^{-x} + 1.20e^{-2x} \tag{6.27}$$

for $0 \leq x \leq 4$. The exponential terms in Equation 6.25 are close to the 'correct' values of Equation 6.27, but the exponents are actually complex in Equation 6.26. Thus, if 'curve-fitting' (Chapter 1) is the only objective and not determination of the parameters themselves, then Equations 6.25 and 6.26 are equally satisfactory.

Equation 6.26 illustrates another problem with linear compartmental models in that it is often difficult to ensure numerically that all k_j are real. In Prony's method, since the k_j are determined from the zeros of a polynomial, they may in general turn out to be complex.

Since Equation 6.20 results in a separable least-squares problem (Section 6.1), a better alternative to Prony's method is probably to treat it as such. The procedure of Wiscombe and Evans (1977) takes this viewpoint and apparently overcomes all *numerical* difficulties due to ill-conditioning. We strongly recommend this paper (a computer program can apparently be obtained by writing to the authors), especially when J in Equation 6.20 is larger than about 2. These authors do not, however, determine confidence

regions for the parameters; these can be readily determined using the methods of Chapter 4 of this manual.

6.4 Independent Variables Subject to Error

We treat here a simple case with one independent variable and one dependent variable. More complicated situations are discussed in Bard (1974). See also Tukey (1975).

If the true values of x_i and y_i are x_i^* and y_i^* respectively, then the mathematical model is

$$y_i^* - f(x_i^*, \underline{p}) = 0; \quad i=1,2,\ldots,n. \tag{6.28}$$

If the statistical models for both x and y are of Type A (and the measurement errors are independent), then the likelihood function is given by (apart from constant terms)

$$-\log L(x^*, y^*, p, \sigma_x, \sigma_y) = \frac{n}{2} \log \sigma_x^2 + \frac{n}{2} \log \sigma_y^2 + \frac{1}{2\sigma_x^2} \sum_{i=1}^{n} (x_i - x_i^*)^2 + \frac{1}{2\sigma_y^2} \sum_{i=1}^{n} (y_i - y_i^*)^2 \tag{6.29}$$

where x_i^* and y_i^* are the 'true' values of x_i and y_i. To proceed further, we must know either σ_x or σ_y (Bard, 1974, p. 68). Then the remaining quantity may be determined by maximizing L. For example, suppose σ_x is known. Then minimizing -log L with respect to σ_y yields

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - y_i^*)^2. \tag{6.30}$$

We now use Equation 6.30 in Equation 6.29 resulting in the constrained minimization problem

$$\min_{\underline{x}^*,\underline{y}^*,\underline{p}} \{\frac{n}{2} \log \sum_{i=1}^{n} (y_i - y_i^*) + \frac{1}{2\sigma_x^2} \sum_{i=1}^{n} (x_i - x_i^*)^2\} \tag{6.31}$$

such that

$$y_i^* - f(x_i^*,\underline{p}) = 0;\ i=1,2,\ldots,n. \tag{6.32}$$

In the case of the model Equation 6.28, the constraints may simply be substituted in Equation 6.31. However, in cases where the model is only given implicitly, the formulation of Equations 6.31 and 6.32 is more convenient (see Section 6.2). Equation 6.32 then becomes

$$\phi(x_i^*,y_i^*,\underline{p}) = 0;\ i=1,2,\ldots,n. \tag{6.33}$$

CHAPTER 7

A WORKED EXAMPLE

In this chapter we present details of some of the calculations discussed in the body of the manual in the context of a simple example.

For the observational data (0., 1.99), (1.00,0.73), (2.00,0.28), (3.00,0.10), it is desired to fit the model

$$y_i = a\, e^{bx_i} + e_i \tag{7.1}$$

where a Type A statistical model is assumed for y_i and x_i is known exactly.

Equation 7.1 is a linear compartmental model (Section 6.3) and yields a separable least-squares problem using the maximum likelihood method (section 6.1). Since there is only one nonlinear parameter (b), the fitting of Equation 7.1 is readily performed by trial and error using a hand-held calculator. For fixed values of $\hat{b}$, $\hat{a}$ may be determined by linear least-squares procedures as

$$\hat{a} = \sum_{i=1}^{4} X_i\, y_i / \sum_i X_i^2 \tag{7.2}$$

where

$$X_i = e^{\hat{b}x_i}. \tag{7.3}$$

The objective function to be minimized in fitting Equation 7.1 is

$$S(\hat{a},\hat{b}) = \sum_{i=1}^{4} (y_i - \hat{a}\, e^{\hat{b}x_i})^2 \tag{7.4}$$

This becomes, on using Equation 7.2,

$$S(\hat{b}) = \tfrac{1}{2}\left[\sum_i y_i^2 - \left(\sum_i x_i y_i\right)^2 / \sum_i x_i^2\right]. \qquad (7.5)$$

In Table 7.1, we show some values of $\hat{b}$ and the corresponding $\hat{a}$ and $S(\hat{b})$ resulting from Equations 7.2 and 7.5.

Table 7.1. Values of $\hat{a}$, $\hat{b}$, and $S(\hat{b})$ from Equations 7.2 and 7.5

$\hat{b}$	$S(\hat{b}) \times 10^5$	$\hat{a}$
-1.00	5.94	1.991
-0.98	13.7	1.984
-0.99	5.69	1.988
-0.994	4.82	1.9887
-0.995	4.80	1.9891
-0.996	4.87	1.9894

The minimum sum of squares occurs at $\hat{b} = -0.995$, $\hat{a} = 1.989$. From Equation 2.19, the estimate of the variance is $\hat{\sigma}^2 = 4.797 \times 10^{-5}$. The data and the fitted curve are shown in Figure 7.1.

To determine approximate joint confidence regions for the parameters, we can use either Equation 4.23 or Equation 4.29. In the former case, the matrix $\underline{B}^T\underline{B}$ is given by

$$\underline{B}^T\underline{B} = \begin{bmatrix} 1.15804 & 0.36174 \\ 0.36174 & 0.92826 \end{bmatrix} \qquad (7.6)$$

The sampling distribution covariance matrix is thus given approximately from Table 4.1 by

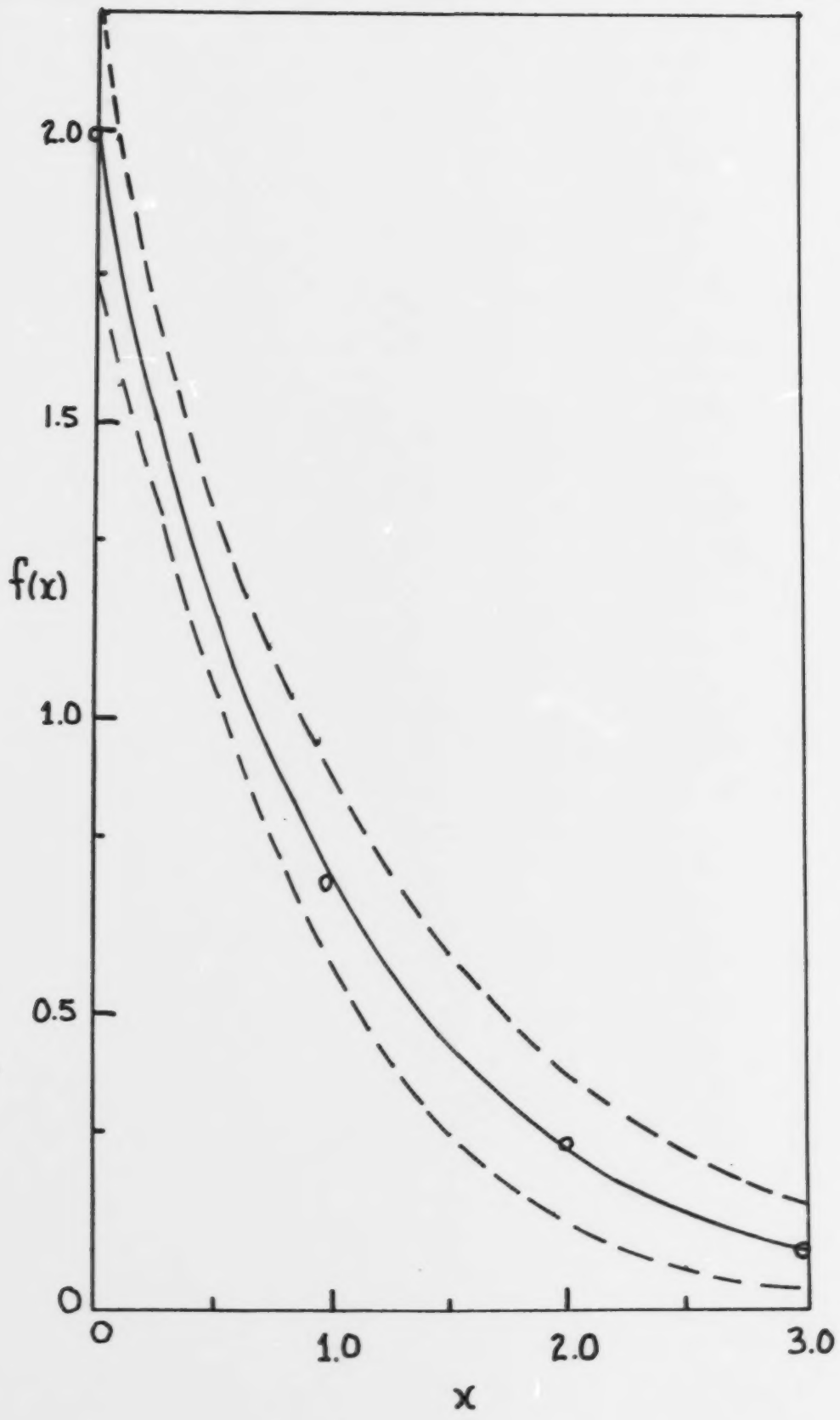

Fig. 7.1. A plot of $f(x) = a\, e^{bx}$ (solid line) and the 99.9 per cent confidence bands for the prediction error (see text). The circles represent the observational data.

$$\underline{V}_p = \hat{\sigma}^2(\underline{B}^T\underline{B})^{-1} = 4.797 \times 10^{-4}\begin{pmatrix} 0.983215 & -0.383156 \\ -0.383156 & 1.226600 \end{pmatrix} \tag{7.7}$$

$$= \begin{pmatrix} 4.716 \times 10^{-5} & -1.838 \times 10^{-5} \\ -1.838 \times 10^{-5} & 5.884 \times 10^{-5} \end{pmatrix}.$$

Equation 4.36 yields the 95% confidence intervals

$$a = 1.989 \pm t_{0.025}\ (4.716 \times 10^{-5})^{\frac{1}{2}} = 1.989 \pm 0.030$$

$$b = -0.995 \pm t_{0.025}\ (5.884 \times 10^{-5})^{\frac{1}{2}} = -0.995 \pm 0.033. \tag{7.8}$$

The Hessian matrix for Equation 7.4 at the minimum is

$$\underline{H}^* = \begin{pmatrix} 1.15804 & 0.36173 \\ 0.36173 & 0.92428 \end{pmatrix}. \tag{7.9}$$

Thus, we see that the approximation of Equation 4.16, which leads to the use of Equation 7.7, is quite good. The upper 5% point of the $F_{2,2}$ distribution is 19.000, and hence a 95% joint confidence interval is given by Equation 4.23, which in the present case is

$$1.15804\ \delta a^2 + 0.72348\ \delta a \delta b\ 0.92827\ \delta b^2 \leq 1.823 \times 10^{-3} \tag{7.10}$$

where

$$\delta a = \hat{a} - 1.989$$
$$\delta b = \hat{b} + 0.995\ . \tag{7.11}$$

The principal components (Section 4.4) are given by

$$\begin{pmatrix} \hat{\rho}_1 \\ \hat{\rho}_1 \end{pmatrix} = \underline{U}^T\,\underline{\hat{\rho}} = \begin{pmatrix} 0.59047 & -0.80708 \\ 0.80708 & 0.59047 \end{pmatrix}\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} 1.97724 \\ 1.01793 \end{pmatrix} \tag{7.12}$$

with covariance matrix

$$\underline{V}_{\rho} = \begin{pmatrix} 0.66361 & 0. \\ 0. & 1.42270 \end{pmatrix} \tag{7.13}$$

and Equation 7.10, when written in terms of the principal components, is

$$0.66361\ \delta\rho_1^2 + 1.42270\ \delta\rho_2^2 \leq 0.001823 \tag{7.14}$$

where

$$\delta\rho_1 = \hat{\rho}_1 - 1.97724$$

$$\delta\rho_2 = \hat{\rho}_2 - 1.01793. \tag{7.15}$$

It is convenient to plot the elliptical boundary of the inequality 7.10 by using Equation 7.14 in conjunction with the inverse transformation of Equation 7.12,

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \underline{U}\hat{\rho} = \begin{pmatrix} 0.59047 & 0.80708 \\ -0.80708 & 0.59047 \end{pmatrix} \begin{pmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \end{pmatrix}. \tag{7.16}$$

The 95% confidence ellipse is shown in Fig. 7.2.

The alternative approach of Equation 4.29 yields the approximate 95% joint confidence region as being given by (a,b) satisfying

$$S(a,b) \leq 20S_0. \tag{7.17}$$

Points on the boundary of the inequality 7.17 are readily determined by trial and error on a hand-held calculator. For the present example, the

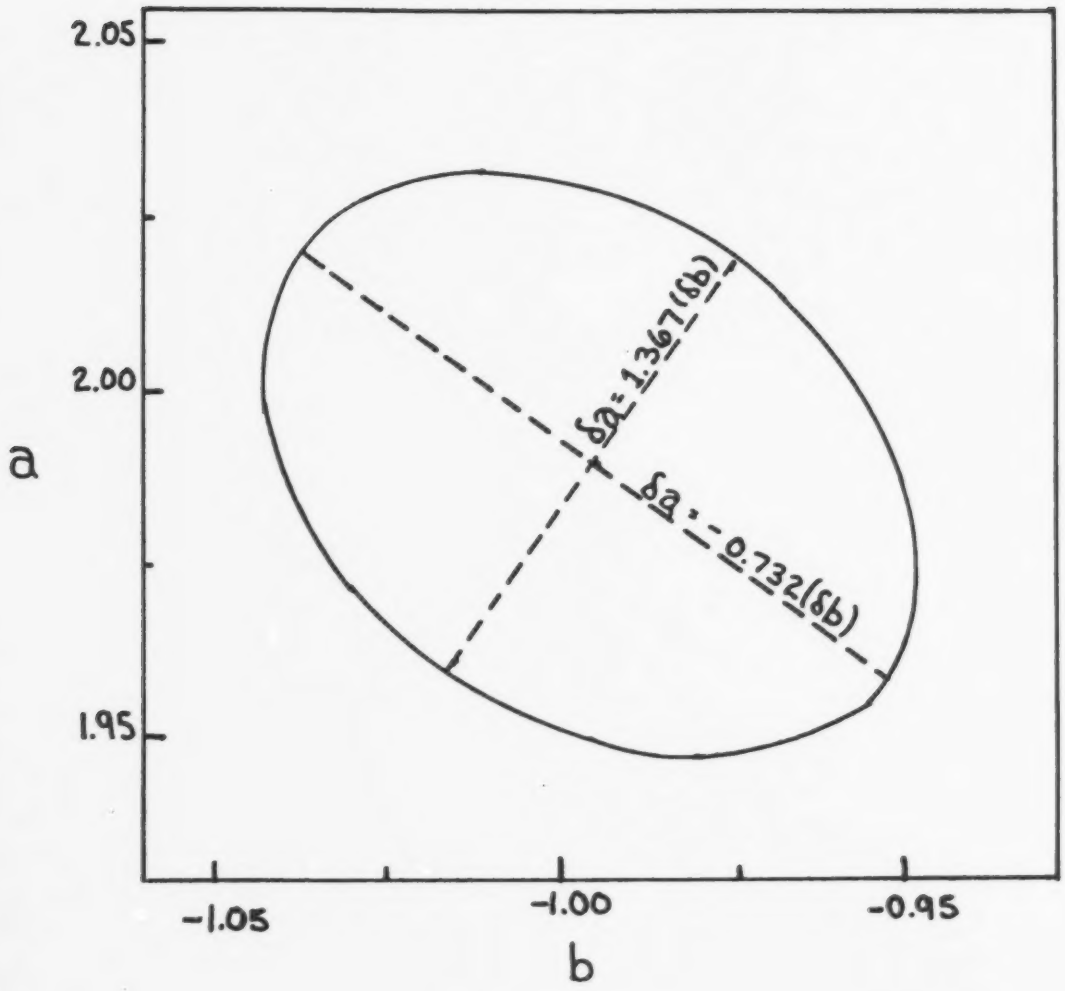

Fig. 7.2 95% joint confidence region, Eqn. (7.10) and Eqn. (7.17), as given by the interior of the ellipse in the figure. The dashed lines indicate the eigenvectors of $\underline{V}_p$ (the columns of $\underline{U}$ in Eqn. (7.12)), and indicate the principal component directions.

results of Equations 7.17 and 7.10 are indistinguishable on the graph of Figure 7.2. This indicates that the approximations used in each case (see Chapter 4) are reasonably accurate.

We note that Equation 7.1 may be linearized by a logarithmic transformation. Using the results of Section 2.4, the errors in $\log y_i$ are approximately normally distributed with variances σ^2/y_i^2. We thus minimize the objective function

$$S(\hat{a}^*,\hat{b}) = \tfrac{1}{2} \sum_i y_i^2 (\log y_i - \hat{a}^* - \hat{b}x_i)^2 \tag{7.18}$$

where $$\hat{a}^* = \log \hat{a}. \tag{7.19}$$

This yields the linear equations

$$\begin{bmatrix} 4.5814 & 0.7197 \\ 0.7197 & 0.9365 \end{bmatrix} \begin{bmatrix} \hat{a}^* \\ \hat{b} \end{bmatrix} = \begin{bmatrix} 2.4399 \\ -0.4364 \end{bmatrix} \tag{7.20}$$

with solution

$$\begin{bmatrix} \hat{a}^* \\ \hat{b} \end{bmatrix} = \begin{bmatrix} 0.68895 \\ -0.99545 \end{bmatrix}. \tag{7.21}$$

Equation 7.21 yields $\hat{a} = 0.992$. We thus see that the parameters determined by minimizing Equation 7.18 are quite similar to those determined by minimizing the full nonlinear sum of squares, Equation 7.3. This has occurred due to the fact that the residuals of the model Equation 7.1 are quite small in the present case, and hence the approximations inherent in using Equation 7.18 are reasonably accurate.

The residuals may be examined (Section 5.1), these being r_i = (0.00101, -0.00559, 0.00795, -0.00062). A plot indicates no apparent inconsistencies with the assumption of the Type A statistical model. We may test the null hypothesis that the true mean residual is zero via the statistic $\sqrt{n}\,\bar{r}/\hat{\sigma}$, which is equal to 0.020 in our case. Since the 5% cutoff points of the Student t distribution with two degrees of freedom are ±4.3027, we have no reason to reject the null hypothesis at the 5% level.

Finally, we can estimate the errors in the predicted true values of y resulting from the mathematical model of Equation 7.1 by using the results of Section 5.2. The variance of the predicted true response at a given x is, from Equation 5.9,

$$E[\delta y(x)^2] = \left(\frac{\partial f}{\partial p}\right)^T \underline{V}_p \left(\frac{\partial f}{\partial p}\right). \tag{7.22}$$

Hence, the standard error of the prediction of Equation 7.1 is given by the square root of the above quantity, which becomes in the present case

$$s_y = e^{bx} (V_{11} + 2axV_{12} + a^2x^2V_{22})^{\frac{1}{2}}. \tag{7.23}$$

The elements V_{ij} of the covariance matrix are given by Equation 7.7. In Figure 7.1, we have shown the 99.9% confidence limits for the true y as given by $y(x) \pm 31.958 s_y$ (31.958 being the symmetric 0.05% points for the Student t distribution with two degrees of freedom).

CHAPTER 8

SUMMARY OF THE PROCEDURE

(1) Establish the mathematical and statistical models (Chapter 1)

$$y_i = f(x_i, \underline{p}) + e_i$$

n measurements y_i at x_i

m parameters $\underline{p}$, where $m < n$

Two common statistical models are

Type A: $e_i \sim N[0, \sigma^2]$

Type B: $e_i \sim N[0, \sigma^2 f^2(x_i, \underline{p})]$

(2) Based on (1), determine appropriate estimator (Chapter 2)

Maximum likelihood gives

$$\text{Type A:} \quad \min_{\underline{p}} \sum_{i=1}^{n} [y_i - f(x_i, \underline{p})]^2$$

$$\text{Type B:} \quad \min_{\underline{p}} \sum_{i=1}^{N} [y_i / f(x_i, \underline{p}) - 1]^2$$

(3) (a) Based on (2), compute estimates $\hat{\underline{p}}$ (Chapter 3)

- Check for special situations:
  - separable problems (Section 6.1)
  - implicit models (Section 6.2)
  - differential equation models (Section 6.2)
  - independent variables subject to error (Section 6.4)
- Also see Bard (1974) and Chambers (1977) for a list of sources of available algorithms.

(b) Compute estimates $\hat{\sigma}$ of σ. For Type A and Type B models, we have

Type A: $\hat{\sigma}^2 = \sum_{i=1}^{n} [y_i - f(x_i, \underline{p})]^2/(n - m)$

Type B: $\hat{\sigma}^2 = \sum_{i-1}^{n} [y_i/f(x_i, \underline{p}) - 1]^2/(n - m)$

(4) Construct approximate joint confidence regions for $\hat{\underline{p}}$ (Chapter 4)

(i) Calculate $\underline{V}_p$ - see Table 4.1 for Type A and Type B models.

If desired, compute confidence intervals for parameters from diagonal entries of $\underline{V}_p$ (Section 4.5).

Consider doing a Monte Carlo calculation of sampling distribution if m not too large (Section 4.6).

(ii) Determine approximate joint confidence regions

Alternative (a):

Type A: $\delta\underline{p}^T(\underline{B}^T\underline{B})\delta\underline{p} \leq m\ \hat{\sigma}^2 F_{m,n-m}$ (Eq. 4.23)

Type B: $\delta\underline{p}^T\{\underline{B}^T \text{diag}[1/f^2(x_i, \underline{p})]\underline{B}\}\delta\underline{p} \leq m\ \hat{\sigma}^2 F_{m,n-m}$ (Eq. 4.24)

Compute principal components of the ellipsoids (Section 4.4)

Alternative (b):

For both Type A and Type B statistical models, determine parameter region satisfying

$$S(\underline{p}) \leq S(\hat{\underline{p}}) \left(1 + \frac{m}{n-m} F_{m,n-m}\right) \quad \text{(Eq. 4.29)}$$

- if m small ($m \leq 3$), alternatives (a) and (b) should both be considered,
- if m large, only alternative (a) is feasible.

(5) Examine residuals (Section 5.1).

(6) Further computations, if desired:

- precision of model predictions(Section 5.2)
- experimental design (Section 5.3)

REFERENCES

Acton, F.S. (1970) *Numerical Methods that Work.* Harper and Row, New York.

Andrews, D.F. (1974) A robust method for linear regression. *Technometrics* 16, 523.

Andrews, D.F. and Pregibon, D. (1978) Finding the outliers that matter. *J. Roy. Stat. Soc.* B 40, No. 1.

Atkinson, A.C. and Hunter, W.G. (1968) The design of experiments for parameter estimation. *Technometrics* 10, 271.

Bard, Y.A. (1974) *Nonlinear Parameter Estimation.* Ac. Press, New York.

Barham, R.H. and Drane, W. (1972) An algorithm for least squares estimation of nonlinear parameters when some of the parameters are linear. *Technometrics* 14, 757.

Beale, E.M.L. (1960) Confidence regions in nonlinear estimation. *J. Roy Stat. Soc.* 22, 41.

Beauchamp, J.J. and Cornell, R.G. (1968) Simultaneous estimation by partial totals for compartmental models. *J. Am. Stat. Ass.*, 573.

Beauchamp, J.J. and Cornell, R.G. (1969) Spearman simultaneous estimation for a compartmental model. *Technometrics* 11, 551.

Betts, J.T. (1976) Solving the nonlinear least square problem: Application of a general method. *J. Opt. Theory & Appl.* 18, 469.

Boole, G. (1960) *A Treatise on the Calculus of Finite Differences.* Dover, New York.

Box, G.E.P. and Draper, N.R. (1965) The Bayesian estimation of common parameters from several responses. *Biometrika* 52, 355.

Box, G.E.P. and Lucas, H.L. (1959) Design of experiments in nonlinear situations. *Biometrika* 46, 77.

Box, M.J. (1968) The occurrence of replications in optimal designs of experiments to estimate parameters in non-linear models. *J. Roy. Stat. Soc.* B 30, 290.

Box, M.J. (1970) Some experiences with a nonlinear experimental design criterion. *Technometrics* 12, 569.

Box, M.J. (1971) Bias in nonlinear estimation. *J. Roy. Stat. Soc.* B 33, 171.

Box, M.J. and Draper, N.R. (1972) Estimation and design criteria for multiresponse non-linear models with non-homogeneous variances. *Appl. Stat.* 21, 13.

Branin, F.H., Jr. (1972) Widely convergent method for finding multiple solutions of simultaneous nonlinear equations. *IBM J. Res. & Dev.* 16, 504.

Branson, D.R., Blau, G.E., Alexander, H.C., and Neely, W.B. (1975) Bioconcentration of 2,2',4,4'-tetrachlorobiphenyl in rainbow trout as measured by an accelerated test. *Trans. Am. Fish. Soc.* 104, 785.

Britt, H.I. and Luecke, R.H. (1973) The estimation of parameters in nonlinear, implicit models. *Technometrics* 15, 233.

Broyden, C.G. (1967) Quasi-Newton methods and their application to function minimization. *Math. Comp.* 21, 368.

Bush, S.F. and Dyer, P. (1976) The experimental and computational determination of complex chemical kinetics mechanisms. *Proc. Roy. Soc. Lond. A* 351, 33.

Cannon, J.R. and Filmer, D.L. (1968) A numerical experiment on the determination of unknown parameters in an analytic system of ordinary differential equations. *Math. Biosciences* 3, 267.

Chambers, J.M. (1973) Fitting nonlinear models: numerical techniques. *Biometrika* 60, 1.

Chambers, J.M. (1977) *Computational Methods for Data Analysis.* Wiley-Interscience, New York.

Chambers, J.M. and Ertel, J.E. (1974) Remark AS R11. *Appl. Stat.* 23, 250.

Cukier, R.I., Levine, H.B., and Shuler, K.E. (1978) Nonlinear sensitivity analysis of multiparameter model systems. *J. Comput. Phys.* 26, 1.

Curtiss, C.F. and Hirschfelder, J.O. (1952) Integration of stiff equations. *Proc. Nat. Acad. Sci. USA* 38, 235.

Cvetanovic, R.J. and Singleton, D.L. (1977) Comment on the evaluation of the Arrhenius parameters by the least squares method. *Int. J. Chem. Kin.* 9, 481.

Davidon, W.C. (1976) New least-squares algorithms. *J. Opt. Theory & Appl.* 18, 187.

Dennis, J.E., Jr. (1977) Nonlinear least-squares and equations. In *The State of the Art in Numerical Analysis,* ed. D. Jacobs, Ac. Press, London.

Dennis, J.E., Jr. and Moré, (1977) Quasi-Newton methods, motivation and theory. *SIAM Rev.* 19, 46.

Dickinson, R.P. and Gelinas, R.J. (1976) Sensitivity analysis of ordinary differential equation systems - a direct method. *J. Comput. Phys.* 21, 123.

Draper, N.R. and Hunter, W.G. (1967) The use of prior distributions in the design of experiments for parameter estimation in non-linear situations. *Biometrika* 54, 147.

Draper, N. and Smith, H. (1966) *Applied Regression Analysis.* Wiley, New York.

Dreyfus, S.E. and Law, A.M. (1977) *The Art and Theory of Dynamic Programming.* Academic Press, New York.

Endrenyi, L. (1976) Statistical problems of kinetic model building. In *Mathematical Models of Metabolic Regulation,* ed. T. Keleti and S. Lakatos, Ac. Press, Budapest.

Endrenyi, L. and Kwong, F.H.F. (1972) In *Analysis and Simulation of Biochemical Systems,* ed. H.C. Hemker and B. Hess, North-Holland, Amsterdam, p. 219.

Gallant, A.R. (1975) Nonlinear regression. *Am. Stat.* 29, 73.

Gauss, K.F. (1809) Theoria Motus Corporum Coelestium. In *Werke,* Vol. 7, 240.

Gear, C.W. (1968) In *Information Processing (IFIP Confernce, Edinburgh),* ed. A.J.H. Morrell, North Holland, Amsterdam, pp. 187-193.

Gear, C.W. (1971) *Numerical Initial Value Problems in Ordinary Differential Equations.* Prentice-Hall, Englewood Cliffs, N.J.

Gill, P.E. and Murray, W. (1978) Algorithms for the solution of the nonlinear least-squares problem. *SIAM J. Num. Anal.* 15, 977.

Goh, B.S. (1978) Global convergence of some differential equation algorithms for solving equations involving positive variables. *BIT* 18, 84.

Goh, B.S. (1979) Differential equation algorithms for minimizing a function subject to nonnegative constraints. In *Applied Nonlinear Analysis,* ed. V. Lakshmikantham, Ac. Press, New York, in press.

Guttman, I. and Meeter, D.A. (1965) On Beale's measures of non-linearity. *Technometrics* 7, 623.

Guttman, I., Pereyra, V. and Scolnik, H.D. (1973) Least squares estimation for a class of non-linear models. *Technometrics* 15, 209.

Hampel, F.R. (1973) Robust estimation: a condensed partial survey. *Z. Warsch.* 27, 87.

Hartley, H.O. (1961) The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares. *Technometrics* 3, 269.

Hartley, H.O. and Booker, A. (1965) Nonlinear least squares estimation. *Ann. Math. Stat.* 36, 638.

Heineken, F.G., Tsuchiya, H.M., and Aris, R. (1967) On the accuracy of determining rate constants in enzymatic reactions. *Biosciences* 1, 115.

Hildebrand, F.B. (1974) *Introduction to Numerical Analysis.* 2nd ed., McGraw-Hill, New York.

Hill, W.J., Hunter, W.G., and Wichern, D.W. (1968) A joint design criterion for the dual problem of model discrimination and parameter estimation. *Technometrics* 10, 145.

Hirsch, M. and Smale, S. (1974) *Differential Equations, Dynamical Systems, and Linear Algebra.* Ac. Press, New York.

Hooke, R. and Jeeves, T.A. (1961) Direct search solution of numerical and statistical problems. *J. Assoc. Comp. Mach.* 8, 212.

Huber, P.J. (1973) Robust regression: asymptotics, conjectures, and Monte Carlo. *Ann. Stat.* 1, 799.

Jakola, T.H.I. and Luus, R. (1974) A note on the application of nonlinear programming to chemical-process optimization. *Opns. Res.* 22, 415.

Jennrich, R.I. (1969) Asymptotic properties of non-linear least squares estimators. *Ann. Math. Stat.* 40, 633.

Jennrich, R.I. and Bright, P.B. (1976) Fitting systems of linear differential equations using computer generated exact derivatives. *Technometrics* 18, 385.

Jennrich, R.I. and Ralston, M.L. (1979) Fitting nonlinear models to data. *Ann. Rev. Biophys. & Bioeng.*, in press.

Jensen, J.A. and Rowland, J.H. (1975) *Methods of Computation.* Scott, Foresman & Co., Glenview, Ill.

Kaufman, L. and Pereyra, V. (1978) A method for separable nonlinear least squares problems with separable nonlinear equality constraints. *SIAM J. Num. Anal.* 15, 12.

Kreider, D.L., Kuller, R.G., Ostberg, D.R., and Perkins, F.W. (1966) *An Introduction to Linear Analysis.* Addison-Wesley, Reading, Mass.

Lawson, C.L. and Hanson, R.J. (1974) *Solving Least Squares Problems.* Prentice-Hall, Englewood Cliffs, N.J.

Lawton, W.H. and Sylvestre, E.A. (1971) Elimination of linear parameters in nonlinear regression. *Technometrics* 13, 461.

Lide, D.R. and Paul, M.S. (eds.) (1974) *Critical Evaluation of Chemical and Physical Structural Information.* National Academy of Sciences, Washington.

Lipton, S. and McGilchrist, C. (1963) Maximum likelihood estimators of parameters in double exponential regression. *Biometrics* 19, 144.

Luus, R. and Jakola, T.H.I. (1973) Optimization by direct search and systematic reduction of the size of search region. *A. I. Ch. E. J.* 19, 760.

Marquardt, D.W. (1963) An algorithm for least-squares estimation of nonlinear parameters. *SIAM J.* 11, 431.

McGilchrist, C.A. (1968) Efficient difference equation estimators in exponential regression. *Ann. Math. Stat.* 39, 1938.

Metzler, C. (1971) Usefulness of the two-compartment open model in pharmacokinetics. *J. Am. Stat. Ass.* 66, 49.

Metzler, C.M., Elfring, G.L., and McEwen, A.J. (1974) A package of computer programs for pharmacokinetic modeling. *Biometrics* 30, 562.

Mikhail, E.M. (1976) *Observations and Least Squares.* I.E.P.-Dun-Donelly, New York.

Minder, Ch.E. and McMillan, I. (1977) Estimation of linear compartmental model parameters using marginal likelihood. *Biometrics* 33, 333.

Myhill, J., Wadsworth, G.P., and Brownell, G.L. (1965) Investigation of an operator method in the analysis of biological tracer data. *Biophys. J.* 5, 89.

Nelder, J.A. and Mead, R. (1965) A simplex method for function minimization. *Comp. J.* 7, 308.

Noyes, R.M. (1977) Application of computer simulation to complicated chemical systems. *J. Phys. Chem.* 81, 2315.

O'Neill, R. (1971) Function minimization using a simplex procedure. *Appl. Stat.* 20, 338.

Parkinson, J.M. and Hutchinson, D. (1972) A consideration of non-gradient algorithms for the unconstrained optimization of functions of high dimensionality. In *Numerical Methods for Non-linear Optimization.* Ac. Press, New York, p. 99.

Parsons, D.H. (1968) Biological problems involving sums of exponential functions of time: A mathematical analysis that reduces experimental time. *Biosciences* 2, 123.

Ralston, A. (1965) *A First Course in Numerical Analysis.* McGraw-Hill, New York.

Rey, W.J.J. (1978) *Robust Statistical Methods.* Springer Lect. Notes in Math. #690.

Rice, J.R. and White, J.S. (1964) Norms for smoothing and estimation. *Siam Rev.* 6, 243.

Saleh, A.K.M.E. and Choudhry, G.H. (1974) *On Fitting Exponential Regressions.* Carleton Univ. Math. Series No. 102 (preprint).

Schwartz, L.M. and Gelb, R.I. (1978) Alternative method of analyzing first-order kinetic data. *Anal. Chem.* 50, 1592.

Seinfeld, J.H. and Lapidus, L. (1974) *Mathematical Methods in Chemical Engineering,* Vol. III: *Process Modeling, Estimation, and Identification.* Prentice-Hall, Englewood Cliffs, N.J.

Sheppard, C.W. (1962) *Basic Principles of the Tracer Method.* Wiley, New York.

Silvert, W.L. (1979) Practical curve fitting. *Limnol. Oceanogr.*, 24, 767.

Sorgan, J.D. and Mikhail, W.M. (1971) A general approximation to the distribution of instrumental variables estimates. *Econometrica* 39, 131.

Spendley, W., Hext, G.R., and Himsworth, F.R. (1962) Sequential applications of simplex designs in optimization and evolutionary operation. *Technometrics* 4, 441.

Swartz, J. and Bremermann, H. (1975) Discussion of parameter estimation in biological modelling: algorithms for estimation and evaluation of the parameters. *J. Math. Biol.* 1, 241.

Tukey, J.W. (1975) Instead of Gauss-Markov least squares, what? In *Applied Statistics,* ed. R.P. Gupta, North-Holland, Amsterdam, p. 351.

Villiers, N. and Glasser, D. (1977) A parameter variation method for non-linear regression. A paper presented at the 70th Annual AIChE meeting, New York, November 1977 (preprint).

Walsh, G.R. (1975) *Methods of Optimization.* John Wiley & Sons, London.

Warner, D.D. (1977) The numerical solution of the equations of chemical kinetics. *J. Phys. Chem.* 81, 2329.

Wiscombe, W.J. and Evans, J.W. (1977) Exponential-sum fitting of radiative transmission functions. *J. Comput. Phys.* 24, 416-444.

Wolfe, M.A. (1976) Some methods for least squares estimation. *J. Inst. Math. Applic.* 18, 219.

Wollberg, J.R. (1967) *Prediction Analysis.* D. Van Nostrand, Princeton, N.J.